EFFICIENT SYSTEM DESIGN: STABILITY AND FLEXIBILITY

A Thesis
Presented to
The Academic Faculty

by

Salih Tekin

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
May 2011

EFFICIENT SYSTEM DESIGN: STABILITY AND FLEXIBILITY

Approved by:

Dr. Sigrún Andradóttir, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Hayriye Ayhan
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Douglas Down
Department of Computing and Software
McMaster University

Dr. David Goldsman
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Seong-Hee Kim
School of Industrial and Systems Engineering
Georgia Institute of Technology

Date Approved: 18 January 2011

To my wife and family.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to people without whom this dissertation would not have been completed.

First of all, I would like to express my deepest gratitude to my advisor, Dr. Sigrún Andradóttir, for giving me the opportunity to do this research, and for her motivation, guidance, and support throughout the entire study. I am also deeply thankful to Dr. Douglas Down for his invaluable comments. Their knowledge, experience, and insights have been very influential and helpful in my studies. I want to thank, Dr. Hayriye Ayhan, Dr. David Goldsman, and Dr. Seong-Hee Kim for their willingness to serve on my thesis committee and for their helpful comments.

Secondly, special thanks go to my parents, Ahmet and Melek, and siblings Büşra and Esra, for all the sacrifices they have made in the past for me to realize my dream.

Last but not least, I would like to express my deepest love and gratitude to my wife, Tuğba, for her continuous support and encouragement. She deserves as much credit as I do for completing this work. I am grateful for her being a part of my life and being by my side to share my happiness and sorrow. I should also mention the joy and happiness that our little one, Ahmet Berke, brought into our lives. This thesis is dedicated to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis is concerned with queueing models where demand is allowed to exceed the system capacity, and also with the capacity sizing and pricing problem for heterogenous products and resources under demand uncertainty. Our aim is to improve productivity and profitability.

In the first part of the thesis, we consider the dynamic assignment of servers to tasks in queueing networks where demand may exceed the capacity for service. The objective is to maximize the system throughput. We use fluid limit analysis to show that several quantities of interest, namely the maximum possible throughput, the maximum throughput for a given arrival rate, the minimum arrival rate that will yield a desired feasible throughput, and the optimal allocations of servers to classes for a given arrival rate and desired throughput, can be computed by solving linear programming problems. We develop generalized round robin policies for assigning servers to classes for a given arrival rate and desired throughput, and show that our policies achieve the desired throughput as long as this throughput is feasible for the arrival rate. We conclude with numerical examples that illustrate the points discussed and provide insights into the system behavior when the arrival rate deviates from the one the system is designed for.

In the second part of the thesis, we consider the effects of inspection and repair stations on the production capacity and product quality in a serial line with possible inspection and repair following each operation. We consider multiple defect types and allow for possible inspection errors that are defect dependent. We construct a profit function that takes into account inspection, repair, and goodwill costs, as well as the

capacity of each station. Then we compare the profitability of different inspection plans and discuss how to identify the optimal inspection plan. Unlike previous works, our analysis captures the possibility of increasing production capacity by scrapping or repairing defective items before a bottleneck operation station, and hence reducing the waste of operation capacity on defective products.

Finally, in the third part of the thesis, we consider the capacity and pricing decisions made by a monopolistic firm producing two heterogenous products under demand uncertainty. The objective is to maximize profit. Our model incorporates dedicated and flexible resources, product substitutability, and processing rates that may depend on the product and on the resource type. We provide the optimum prices and production quantities as functions of resource capacities and demand intercepts. We also show that investment in flexible capacity is only desirable when it is optimal to invest in dedicated capacities for both products, and obtain upper bounds for the costs of the dedicated capacities that need to be satisfied for investment in the flexible resource. We conclude with numerical examples that illustrate the points discussed and provide insights into how the optimal capacities and expected production quantities, prices, and profit depend on various model parameters.

CHAPTER I

INTRODUCTION

In today's highly competitive market, it is vital to find new ways to utilize the existing resources in a production/queueing system more efficiently. This thesis is concerned with the analysis and implications of allowing instability in queueing systems, specifically in serial inspection systems, as well as with capacity and pricing decisions for flexible resources. Our primary objectives include increasing the production rate and profitability of the systems under consideration.

Consider a network with K service facilities (or stations) and M servers assigned to those stations, with probabilistic routing among stations. In traditional queueing network models, each server is dedicated to work only at a single station. However, it is interesting to study the effects of flexible (cross-trained) servers that are capable of working at different stations, with the objective of achieving more efficiency. This interest in flexible workforce has motivated many researchers to determine ways to utilize cross-trained servers efficiently. For example, researchers have considered how servers should be moved dynamically between stations in order to enhance system performance. In analyzing these flexible systems, as well as other queueing systems, the stability of the queue lengths is an implicitly required assumption or goal. There exists only a limited amount of research on unstable queueing systems (see Chapter 2 for a literature review). However, in certain settings, allowing instability can lead to performance improvements.

In the first part of this thesis, we investigate multi-class, discrete-flow networks with infinite buffers when demand is allowed to exceed the capacity for service. Multiple types of customers are serviced by flexible servers that are able to work on several different classes. Offered demand to each class can come from both external sources as well as internal transitions. The same server can have different service rates for different classes. Moving a server among the classes is assumed to incur switching times that can be different for each origin-destination pair of classes. More than one server can be assigned to a given class, possibly with different service rates. In that case, servers at a class can either cooperate by working simultaneously on a customer, or work in parallel and process the customers separately. We concentrate on the case where the servers work in parallel and there is one arrival stream routed

to various classes (cooperating servers and multiple arrival streams are straightforward extensions). Our aim is to find the best assignment of servers to classes so that the throughput of the system is maximized in the multi-class network with flexible servers.

To motivate our analysis, consider manufacturing processes where demand exceeds the production capacity and work in process can be either salvaged for some profit or scrapped at small cost compared to the final product value. In these cases, allowing instability in the system might be desirable given the right parameters. We quantify the effects of allowing instability on both throughput and server assignments, and also construct a linear program that is used to identify the optimal allocation of servers to classes, as well as the resulting throughput. Two server allocation policies are introduced with proofs that they can achieve any throughput less than the optimal value. Through numerical studies, we show how the assignments are determined for a specific network, and provide information about the sensitivity of the optimal assignment with respect to the demand, as well as some simulation results.

Quality and cost are important factors impacting the profitability of competitive manufacturing industries. To prevent nonconforming items from reaching customers, inspection of products is performed. Although repeated inspection may add to the total cost of produced units, it introduces the opportunity of scrapping defective units early in the production process, so as to avoid wasting production capacity, particularly at bottleneck stations, on defective units, as well as eliminating unnecessary production cost incurred for defective units. Therefore, there is a tradeoff between inspection, repair, and scrap costs on the one hand, and the cost of products with undetected defects reaching customers on the other hand.

In the second part of the thesis, we analyze inspection policies for serial production systems based on the total profit rate, where production may be constrained by the external arrival rate (demand), or by the capacity of any of the inspection, repair, or production processes. In addition to factoring in revenue and production, inspection, and repair costs, we also take into account goodwill cost that is incurred when defective units are shipped to customers. Goodwill cost may be incurred directly in the form of repair costs, or indirectly in the form of loss of customer goodwill.

Although previous works take into account the throughput of the system in the inspection allocation problem, they only consider reduction in the throughput as a result of scraps at inspection stations or consider inspection stations that determine the overall throughput (see Chapter 2 for a literature review). None of them quantify the benefit of having inspection, and hence scraps, before the bottleneck stations, which

has the effect of increasing the capacity of the bottleneck. Moreover, our model allows for more generality in the repair and inspection processes. For instance, inspection errors can depend both on the station and defect type; partial inspection is allowed as opposed to complete inspection; at each inspection station, some units are identified as having repairable or not repairable defects, and hence scraps are possible after the inspection stations; and also repair cost and probability at each repair station can depend on whether a unit is actually defective or wrongly classified to be defective. Since we take into account the side effects of inspection on downstream bottleneck stations, our results are more broadly applicable than previous work considering demand constrained systems. We also demonstrate through numerical examples that bottleneck considerations for determining the best inspection locations can lead to different inspection decisions than previous models (that do not take the capacity of the system into account).

The capacity investment decision has a high impact on a firm's profitability and competitiveness. Capacity levels are typically determined in advance of the actual production because of the long lead times for acquiring the capacity. This decision can be made as early as five years before the planned production date (see Fleischmann, and Henrich [39]). Hence, capacity decisions have to be made under demand uncertainty based on available forecasts, resulting in mismatches between supply and demand. As a result, firms are increasingly resorting to flexibility, both on the supply and demand sides, to effectively match their supply with demand (see e.g., Boudette [19], Edmondson [35], Holweg and Pil [53], Jordan and Graves [58], Mackintosh [68], McMurray [70], and Muriel, Somasundaram, and Zhang [71] for specific examples).

In the third part of the thesis, we analyze the optimal capacity decision faced by a price-setting and monopolistic firm producing two substitutable or complementary products with flexible and dedicated technologies. The firm needs to determine its production capacity beforehand under demand uncertainty (first stage); however, production and pricing decisions can be made after the demand uncertainty is resolved (second stage). We assume a linear demand function, where demand for a product is inversely related to its own price, taking into account possible cross-price effects on demand from the other product. Demand uncertainty is introduced into the model as uncertainty about the location of demand curves. There are two types of capacity, namely dedicated to one product and flexible to produce both products. We allow products to be heterogenous in that they might require different amounts of server time. Also, the flexible server may have service rates that differ from those of the dedicated servers. Determining how to share the flexible capacity between the

two products is another decision variable in the second stage when actual demand parameters become visible. Hence the firm can deal with the demand uncertainty by changing the supplies for both products through flexible capacity (supply side flexibility), and by changing the demand for the products through pricing (demand side flexibility). The firm can also produce below the installed capacity (volume flexibility). Although previous works consider similar two stage problems, none of them model the effects of server capabilities, along with product substitutability and different technology types (dedicated and flexible), on the optimal capacity decision, as well as the optimal expected profits. Hence, our model allows for more generality and broader applicability.

The remainder of this thesis is organized as follows. In Chapter 2, we review the previous research on flexible servers, inspection allocation, and capacity sizing and pricing. In Chapter 3, we provide our analysis and results on dynamic server allocation with flexible servers. In Chapter 4, we describe the inspection allocation problem in capacity-constrained serial lines and provide our results. We discuss and analyze the capacity planning and pricing problem in Chapter 5. Finally, we summarize our main results and contributions and suggest possible future research directions in Chapter 6.

CHAPTER II

LITERATURE REVIEW

In this chapter, we review the literature that is related to our work. More specifically, Section 2.1 reviews the literature on server assignment to queues (both static and dynamic) for increasing the efficiency of flow lines, as well as results on unstable queues. Next, we review the research on the inspection allocation problem specifically for serial network configurations in Section 2.2. Finally, Section 2.3 reviews the literature on pricing and capacity allocation problems.

2.1 Flexible Server and Unstable Queueing Literature

In recent years, there has been a growing interest in queueing systems with flexible servers, with most of the work examining holding costs or throughput. Ahn, Duenyas, and Zhang [2], Pandelis and Teneketzis [76], and Farrar [37] study how servers should be assigned to stations to minimize the total holding cost incurred for systems with two queues in tandem and no arrivals. Ahn, Duenyas, and Zhang [3] consider the same problem for a two-class queueing system with one dedicated server, one flexible server, and no exogenous arrivals. Similarly, for systems with two stations and Poisson arrivals, Ahn, Duenyas, and Lewis [1], Hajek [50], and Rosberg, Varaiya, and Walrand [82] aim to minimize the expected total holding cost by assigning servers to stations. Works that aim to maximize the long-run average throughput through dynamic assignment of reliable servers include Andradóttir, Ayhan, and Down [4, 5, 6] and Tassiulas et al. [87, 88]. By contrast, Andradóttir, Ayhan, and Down [7, 8] and Wu, Lewis, and Veatch [98] determine the optimal allocation of flexible servers in a tandem-line system where servers are not necessarily reliable. For parallel queueing systems with flexible servers and external arrivals, Williams [97], Bell and Williams [10, 11], Bramson and Williams [20], and Harrison and López [52] suggest asymptotically optimal server assignment policies that minimize the discounted infinite-horizon holding costs under a heavy traffic assumption.

The earliest work we are aware of that considers overloaded systems is by Goodman and Massey [43]. They study non-ergodic Jackson networks and propose a way to determine the maximal subnetwork that achieves steady state. Weiss [95] considers a Jackson network in which some of the nodes have an infinite supply of customers. He

shows that when only customers in transit are counted as congestion, the stable subset of nodes has the usual product-form distribution. Similarly, the marginal distribution for the number of customers in transit exists for each node with an infinite supply of work, but the joint distribution does not have product-form. Kopzon, Nazarathy, and Weiss [62] and Nazarathy and Weiss [74] determine policies for push-pull networks that ensure that the networks are working at full utilization.

Chen and Mandelbaum [23] conduct a bottleneck analysis of a dynamic, discrete-flow network, where customers are indistinguishable. They use a fluid approximation of the initial discrete network to identify the system throughput, and show that calculating equilibrium throughput rates is equivalent to identifying the bottlenecks of the original network. Unlike our work, in their network, servers are dedicated to a single class. We will find that allowing the servers to be flexible considerably complicates the analysis, as it is difficult to precisely control the amount of time a server spends at each class. A diffusion approximation for the fluid model in Chen and Mandelbaum [23] is described by the same authors in [24]. Andradóttir, Ayhan, and Down [6] identify a tight upper bound on the capacity, while maintaining stability, and provide a method to construct server assignment policies with performance arbitrarily close to this bound. By contrast, we do not require the system to be stable, which also significantly complicates the analysis. Note that if the class of a customer determines the server (that is if only one server is allowed per class) and the servers are not allowed to move, then our problem reduces to production scheduling of classes at each node.

Overloaded systems have also been considered in nonstandard queueing networks where the service rates at the individual classes are not independent, but depend deterministically on the state of the entire system. In such a network, Jonckheere, van der Mei, and van der Weij [57] obtain necessary conditions for rate stability at each class, and also provide bounds for the output rate at each class. Similarly, for bandwidth sharing networks, Egorova, Borst, and Zwart [36] give a partial characterization of the overloaded system's behavior by providing a fixed-point equation for the asymptotic growth rates of the queue lengths.

The use of fluid limits in queueing systems is by now a standard technique. From a stability point of view, it is known that stability of the fluid model is intimately related to the stability of the queueing network (Dai [30]). It is also known that if the fluid model of a queueing network is unstable in a strong sense, then the queueing network is unstable in the sense that the total number of customers in the queueing network diverges (Dai [31]). However, additional analysis is required to address how

a network with flexible servers becomes unstable.

2.2 Inspection Allocation Literature

The inspection allocation problem has been studied by many researchers under different assumptions on the topology of the assembly line. Raz [80] and, more recently, Mandroli et al. [69] provide exhaustive reviews of the work done in this area. Next, we review some works on the inspection allocation problem, namely the early papers and ones closely related to our work.

Lindsay and Bishop [66] proposed a serial production line model where inspection stations are assumed to be error free and all defects are scrapped. Each inspection station could only check the outcome of the previous operation station. They showed that the optimum inspection level at each station was to inspect either all or no units. Hurst [56] was the first to account for the occurrence of inspection errors. Raz and Kaspi [81] examined sequencing and location issues for serial lines with rework and scrap, where multiple inspections are possible after a given operation. Cochran and Erol [29] developed analytical models for calculating the outgoing quality level in a serial manufacturing line with repair stations, scraps, multiple defect types, and inspection errors. Garcia et al. [42], Yum and McDowell [99], and Britney [21] studied the optimal allocation of inspection stations for non-serial production systems, where the output of a processing activity can serve as input to multiple operations. Re-entrant production systems with inspection at various stages of processing were studied by Narahari and Khan [73]. Lee and Unnikrishnan [65] considered a job shop system with finite inspection resources, where each part had a particular manufacturing sequence. More recently, Galante and Passannanti [41] proposed an integrated approach to part scheduling and inspection for job shop manufacturing. Foster et al. [40] and Villalobos et al. [93] discussed the optimal inspection allocation problem for flexible inspection systems where the decisions on what parts to inspect are made just prior to performing inspection operations. Unlike the above studies which considered optimization of steady-state performance, Kogan and Raz [61] studied the problem in a finite planning horizon setting.

Since the advent of continuous improvement methodologies, it has been of interest to know how defective a product is, not just whether it is defective. All of the papers reviewed in the previous paragraph consider workstations of attribute data (WAD), where defects are introduced with Bernoulli type distributions. Some papers proposed models with workstations of variable data (WVD), where defect values have

continuous probability density functions. Hsu and Tapiero [54, 55, 86] constructed and analyzed queue dependent sampling plans intended to screen defective products for single stage M/M/1 and M/G/1 queues based on WVD. Chevalier and Wein [27] proposed a mathematical model based on WVD in a serial line. Their model was the first to address the presence of multiple defect types, and involves the joint optimization of the inspection allocation and testing policy, together with an application involving Hewlett-Packard. However, every defective part is assumed to be repaired with probability one, and hence no scraps exist. Shiau [83, 84] considered inspection capability, manufacturing capability, and tolerance in the inspection allocation problem with the WVD inspection error model. More recently, Volsem et al. [94] developed an evolutionary algorithm that jointly optimizes the number and location of inspection stations, as well as the inspection limits (acceptance range), for a serial multi-stage production line.

The tradeoff in all the above models is between inspection, repair, scrap, and goodwill costs. Such traditional economic models of optimal inspection allocation focus on minimizing the production cost while meeting minimum required product quality levels. However, none of them explicitly accounts for the fact that inspection and repair may place an additional burden on the system, causing productivity to decrease. As a result, the optimal solution takes the form of all or none inspection at each inspection station. Moreover, the works discussed so far all assume infinite buffers between stations. Shin et al. [85] considered the effects of all or none inspection on the throughput in a WAD serial line model with infinite buffers, a single defect type, and the objective to satisfy a throughput requirement. Gurnani et al. [49] constructed a two stage serial line WAD model, with finite buffers among stations, having error free, all or none inspection operations, and compare the case with an inspection station at the end to the case when there are inspection stations after each production operation. Han et al. [51] provided a closed form approximation for the average steady-state production rate in a serial line with all or none inspections, single defect type, and finite buffers between the stages.

Some authors considered the effects of inspection on the production capacity of the system. Bai and Yun [9] discussed the problem where a limited number of inspection stations are available in a serial line with single defect type and perfect repair. However, their model is restricted to the case where production is constrained by the throughput rate of the inspection systems, so that adding new inspection systems or increasing the inspection level (representing the percentage of defects inspected for) might lower the total throughput of the system. Kakade et al. [59] extended

the results of Bai and Yun [9] by accounting for dissimilar quality characteristics carried by different components, with the inspection time depending on component quality. In both models, the inspection level is the same for all inspection stations, and hence all the inspection stations have the same capacity. Rau and Chu [78, 79] accounted for the existence of both types of workstations (WAD and WVD) in arrival-constrained serial lines and re-entrant production systems, respectively, with rework, repair, scrap, and a single defect type. Although both papers determine the inspection allocation based on optimal profit, they only take into account the effect of scrapped units on the throughput. Kim and Gershwin [60] analyzed how production system design, quality, and throughput are interrelated in a small production system with two WVD type machines without scraps and finite buffers. Instead of scrapping defective units, they employed a continuous improvement policy, where the system is stopped once the machine is out of order and producing defective units. Kouikoglou and Phillis [63] jointly determined inspection and production plans for a single stage, input-constrained production system taking into account the throughput rate. More recently, Penn and Raviv [77] suggested a polynomial time algorithm for determining the location of error free, all or none inspection stations in an arrival-constrained serial line with a single defect type and no repair, so that expected net profit per time unit is maximized.

2.3 Capacity Sizing and Pricing Literature

The value of capacity flexibility has been extensively studied in the operations management literature, with most works assuming exogenously given prices. Linear demand models, where demand for a product is inversely related to its price, are very common in the capacity management literature, since the linear form presents a reasonable representation of the demand price relationship while providing analytical tractability (see Bish and Suwandechochai [16]). All of the papers that we mention below consider linear demand models similar to the one considered in this thesis.

Netessine, Dibson, and Shumsky [75] consider a model with partially flexible resources where higher valued resources can supply the demand for any of the lower valued resources. For the case of two products under exogenously given prices, they show that increasing demand correlation causes a shift in capacity choice from flexible to dedicated resources. Van Mieghem [90] finds that flexibility is beneficial even under perfect positive correlation of demand if one product is more profitable than the other for a two product firm facing a bivariate demand distribution. His numerical

results show that as the demand correlation between products increases, the optimal dedicated resource capacity increases in a concave manner, while optimal flexible resource capacity decreases in a convex manner. Later, Van Mieghem [91] compares a flexible resource strategy with a component commonality strategy and shows their equivalence under exogenously given prices.

In all of the papers mentioned in the previous paragraph, prices are exogenously given. Other works consider models where the firm has control over the product pricing, and hence can affect its demand. The ability to modify price before product launch is known as pricing flexibility. Other forms of flexibility include volume and product flexibilities, which imply the ability to produce below installed capacity and switch capacity between products without cost or penalty. Fine and Freund [38] determine the optimal level and mix of dedicated and flexible capacities for a firm manufacturing two products under demand uncertainty, with the demands restricted to only take a discrete set of possible values and the pricing decision implicitly considered through a concave revenue function. They show numerically that the expected profit and total capacity are increasing in demand variance (because the extra revenue when demand and prices are high dominates the loss in revenue when demand and prices are low). Gupta, Gerchak, and Buzacott [48] extend the model of Fine and Freund [38] by studying the effects of existing capacities through numerical examples.

Van Mieghem, and Dada [92] conduct a comprehensive study of postponement (flexibility) strategies for a single product, with the firm deciding on optimal capacity levels, production quantities, and price. Biller, Muriel, and Zhang [13] study the impact of price postponement on capacity and flexibility investment decisions. Through a numerical study, they show that considering price postponement (flexibility) at the planning stage leads to reduction in capacity investments, especially for the flexible capacity, and hence to an increase in profits. The trade-off between dedicated and more expensive flexible resources and the firm's optimal capacity decision has been analytically characterized by Bish and Wang [18] and Bish and Hong [15]. Bish and Wang [18] provide threshold values calculated from the model parameters that can be compared with the unit cost of the flexible resources to determine if the investment in flexible resources is profitable. Bish and Hong [15] consider systems with two products and two resources where the resource that can be used to produce the higher level product also can be used to produce the lower level one, but not vice versa. They obtain the optimal investment strategy when the demands for the products have perfect positive and negative correlation.

Although all of the mentioned research considers investment in a mix of dedicated

and flexible capacities, they assume that products are independent with no cross price effect and that the flexible and dedicated resources produce all the products at the same rate. Some researchers consider general linear demand models with cross-price effects (known as product substitutability/complementarity) and correlation on demand for each product. However, to keep analytical tractability and emphasize their results on substitutability/complementarity, they allowed investment either in dedicated or flexible resources, not both.

In particular, assuming that the random demand intercepts have a normal distribution, Chod and Rudi [28] show that the investment of a monopolist in flexible capacity increases in both demand variability and correlation. However, they do not analyze the sensitivity of capacity decisions with respect to product substitutability. Goyal, Netessine, and Randall [47] also conduct empirical analysis of the North American automotive industry and show that flexible capacity is most valuable with high demand uncertainty but low demand correlation. Goyal and Netessine [44, 45] support the results of Chod and Rudi [28] and further explore the impact of product substitutability on optimal capacity. Birge, Drogosz, and Duenyas [14] and Bish and Suwandechochai [17] also study the impact of substitutability on optimal capacity investment decisions with flexible resources and show that total capacity investment increases with the substitutability parameter. Goyal and Netessine [46] show that while the value of product flexibility always decreases in demand correlation, the value of volume flexibility can increase or decrease in demand correlation depending on whether the products are strategic complements or substitutes. They also show that volume flexibility is a better tool against aggregate demand uncertainty for the two products, while product flexibility is better at mitigating the individual demand uncertainties. More recently, Bish, Liu, and Suwandechochai [16] consider a linear demand model where uncertainty can be included in the model either as an additive or as a multiplicative variable, and study how various market conditions and assumptions on demand (additive or multiplicative) affect a monopolist firm's capacity investment decision under responsive pricing. Finally, Lus and Muriel [67] consider a model with flexible and dedicated resources and product substitutability, and conduct a numerical analysis to study the effects of a substitutability parameter on optimal profits and flexible capacity for a monopolistic firm producing two products and various models where two competing firms each produce one product.

CHAPTER III

DYNAMIC SERVER ALLOCATION FOR UNSTABLE QUEUEING NETWORKS WITH FLEXIBLE SERVERS

In this chapter we study general queueing networks with multiple customer classes and flexible servers without any stability restrictions on the network. Our objective is to construct an assignment policy for the servers such that the throughput from the system is maximized. We will use fluid limit analysis to quantify the effects of allowing instability on both throughput and server assignments.

The organization of this chapter is as follows. In Section 3.1, our queueing network model and other assumptions are described in detail. In Section 3.2, we construct a linear program (LP) that is used to identify the optimal allocation of servers to classes, as well as the resulting throughput, and we provide a uniqueness result for the sets of stable or unstable classes. Section 3.3 introduces two server allocation policies with proofs that they can achieve any throughput less than the optimal value. In Section 3.4, the concepts of "saturation" input and maximum output that can be achieved are introduced, as well as modified linear programs to calculate those quantities. Section 3.5 gives numerical results that show how the assignments are determined for a specific network, and provides information about the sensitivity of the optimal assignment with respect to the demand, as well as some simulation results. Finally, Section 3.6 summarizes our findings.

3.1 Queueing Network Model

We consider a network composed of M flexible servers and K classes of customers, with a buffer of infinite size for each class. The class of a customer represents its current processing stage and customers can change class after each stage. The classes may all be at separate physical stations or there may be several classes served at a particular station. The network is supplied by an exogenous arrival process with independent and identically distributed (i.i.d.) increments $u(n)$ for the n^{th} customer with $E(u(1)) = 1/\lambda$. An external arrival is routed to class k with probability $p_{0,k}$, for $k = 1, \ldots, K$. Let the resulting interarrival time of the nth customer at class k be denoted by $u_k(n)$. We allow $p_{0,k} = 0$ for some k, meaning that the external arrival process for customers to class k is null. The arrival rate to class k is denoted by

$\lambda_k = \lambda p_{0,k}$.

Upon completion of service, a class i customer becomes one of class k with probability $p_{i,k}$, with the customer leaving the system with probability $p_{i,0} = 1 - \sum_{k=1}^{K} p_{i,k}$ for $i, k = 1, \ldots, K$. Let the routing matrix P have (i, k) entry $p_{i,k}$ for $i, k = 1, \ldots, K$. We assume that the n-step transition matrix P^n satisfies $P^n \to 0$ as $n \to \infty$, which implies that the network is an open network and $(I - P)^{-1}$ exists and is nonnegative.

The servers are assumed to be flexible, with each server being capable of serving a set of classes. If server j is capable of serving class k, then the n^{th} customer served by server j at class k has a service time given by $v_{j,k}(n)$, so that the service rate at a class can depend on both the server and the class being served. We assume that the sequence $\{v_{j,k}(n)\}$ is i.i.d. for each $j = 1, \ldots, M$ and $k = 1, \ldots, K$. The mean service time is given by $m_{j,k} = E(v_{j,k}(1))$ for server j at class k, with corresponding service rate $\mu_{j,k} = 1/m_{j,k}$. If server j is not capable of serving class k, we set $v_{j,k}(n) = \infty$ and $\mu_{j,k} = 0$. Within a class, service is First Come First Served (FCFS). Moving server j from class i to class k the nth time incurs a switching time $\xi^j_{i,k}(n)$, $i, k = 1, \ldots, K$, $j = 1, \ldots, M$. We assume that the sequence $\{\xi^j_{i,k}(n)\}$ is i.i.d. for each $i, k = 1, \ldots, K$, $j = 1, \ldots, M$ with mean $s^j_{i,k} = E(\xi^j_{i,k}(1))$. The interarrival, service and switchover times are assumed to be mutually independent.

Next we define some cumulative processes for the queueing network model. The total number of exogenous arrivals at time t is represented by $E_0(t)$. The processes $A = \{A(t), t \geq 0\}$, $E = \{E(t), t \geq 0\}$, and $D = \{D(t), t \geq 0\}$ are K-dimensional column vectors with $A_k(t)$ denoting the cumulative number of class k customers that arrive in $(0, t]$, $E_k(t)$ being the number of exogenous arrivals to class k in $(0, t]$, and $D_k(t)$ being the total number of departures from class k in $(0, t]$. The variable $\Phi_{i,k}(n) = \sum_{l=1}^{n} \phi_{i,k}(l)$, $i = 1, \ldots, N$, $k = 0, \ldots, N$ is the number of customers that arrive to class k from class i among the first n customers passing through class i (with the k=0 case corresponding to departures from the system), and $\phi_{i,k}(n)$ are independent Bernoulli random variables that have value one with probability $p_{i,k}$ (meaning that the n^{th} customer from class i is routed to class k) and are zero otherwise. Moreover, $V_{j,k}(t)$ is the residual service time for class k by server j at time t (set to infinity if $\mu_{j,k} = 0$) and $U(t)$, $U_k(t)$ are the residual exogenous interarrival time at time t to the system and to class k, respectively. Let $T_{j,k}(t)$ be the total amount of time that server j spends serving class k customers in $(0, t]$ and $S_{j,k}(t)$ be the potential number of service completions by server j at class k if server j devotes all its time to class k in $(0, t]$. Finally, let $W^j_{i,k}(n)$ denote the total time spent by server j on switching from class i to class k up to and including the nth switch.

Expressing the cumulative processes in terms of the interarrival, service, and switching times $u_k(n)$, $v_{j,k}(n)$, and $\xi_{i,k}^j(n)$, we have

$$\begin{aligned} S_{j,k}(t) &= \max\{n : V_{j,k}(0) + v_{j,k}(1) + v_{j,k}(2) + \cdots + v_{j,k}(n-1) \le t\}; & (1)\\ E_0(t) &= \max\{n : U(0) + u(1) + u(2) + \cdots + u(n-1) \le t\}; & (2)\\ E_k(t) &= \max\{n : U_k(0) + u_k(1) + u_k(2) + \cdots + u_k(n-1) \le t\}; & (3)\\ W_{i,k}^j(n) &= \sum_{m=1}^{n} \xi_{i,k}^j(m). & (4) \end{aligned}$$

By the Strong Law of Large Numbers (SLLN), we have,

$$\lim_{t\to\infty} \frac{E_k(t)}{t} = \lambda_k, \ \lim_{t\to\infty} \frac{S_{j,k}(t)}{t} = \mu_{j,k}, \text{ and } \lim_{n\to\infty} \frac{W_{i,k}^j(n)}{n} = s_{i,k}^j, \quad \text{for } j = 1, \ldots, M, \text{ and } i, k = 1, \ldots, K. \quad (5)$$

Finally, we assume that the interarrival times are unbounded and spread out. That is, there exists some integer l, and some function $q(x) \ge 0$ on $\mathbb{R}_+$ with $\int_0^\infty q(x)dx > 0$, such that

$$\begin{aligned} P(u(1) \ge x) &> 0, \text{ for any } x > 0, & (6)\\ P(a \le u(1) + \ldots + u(l) \le b) &\ge \int_a^b q(x)dx, \text{ for any } 0 \le a < b. & (7) \end{aligned}$$

This assumption is required for Theorem 4.2 in Dai [30], which we will use in Section 3.3.5.

Let the queue length at class k at time t be denoted by $Q_k(t)$. For a given server assignment policy (i.e., the functions $T_{j,k}(t)$ are given for all j and k), the cumulative variables satisfy the following queueing network equations

$$\begin{aligned} A_k(t) &= E_k(t) + \sum_{i=1}^{K} \Phi_{i,k}(D_i(t)), \ k = 1, \ldots, K; & (8)\\ D_k(t) &= \sum_{j=1}^{M} S_{j,k}(T_{j,k}(t)), \ k = 1, \ldots, K; & (9)\\ Q_k(t) &= Q_k(0) + A_k(t) - D_k(t), \ k = 1, \ldots, K; & (10) \end{aligned}$$

and $0 \le \sum_{k=1}^{K} T_{j,k}(t) \le t, \ j = 1, \ldots, M$. Finally, let $D(t) = \sum_{k=1}^{K} \Phi_{k,0}(D_k(t))$ be the total number of departures from the system until time t. Then the throughput of the system is given by $\limsup_{t\to\infty} D(t)/t$.

3.2 Deterministic Analysis

When we allow instability in the system, the calculation of the flow rates at each class is not obvious. In particular, the usual traffic equation for the flow rate at class k (i.e., $r_k = \lambda p_{0,k} + \sum_{i=1}^{K} p_{i,k} r_i$, where r_k is the effective inflow rate to class k) is not valid, because in our case the input rate to a class does not necessarily equal the output rate from that class. In this section, given the offered demand λ to the system, we construct an optimization problem whose solution provides both the optimal allocation of servers to classes and also the corresponding input and output rates at each class. The allocation of the servers is such that the maximum capacity for the network is achieved for λ, while satisfying network constraints.

The outline of this section is as follows. In Section 3.2.1 the LP that is used to determine the allocation of servers, is constructed. Section 3.2.2 introduces a uniqueness result for the effective inflow and outflow from each node in the network given the allocation parameters. Finally, in Section 3.2.3, we identify the stable and unstable classes based on the allocation LP, and also consider the special case when we have a Jackson network.

3.2.1 The Allocation LP

In this section, we introduce the allocation LP. We start by defining the flows within the network. The effective inflow rate a_k to class k consists of inflow from the outside plus the inflow from the other classes within the network. Similarly, d_k is the effective outflow rate from class k. Let $\delta_{j,k}$ be the fraction of time that server j devotes for class k customers. For all $k = 1, \ldots, K$, we have

$$a_k = \lambda p_{0,k} + \sum_{i=1}^{K} d_i p_{i,k}, \tag{11}$$

$$d_k = \min\Big(\sum_{j=1}^{M} \mu_{j,k} \delta_{j,k}, a_k \Big). \tag{12}$$

Next, we maximize the throughput using the following allocation LP having decision variables d_k and $\delta_{j,k}$ for $j = 1, \ldots, M$, $k = 1, \ldots, K$:

$$\max \sum_{k=1}^{K} d_k p_{k,0} \text{ such that} \tag{13}$$

$$d_k \leq \sum_{j=1}^{M} \mu_{j,k}\delta_{j,k}, \ k = 1, \ldots, K; \tag{14}$$

$$d_k \leq \lambda p_{0,k} + \sum_{i=1}^{K} d_i p_{i,k}, \ k = 1, \ldots, K; \tag{15}$$

$$\sum_{k=1}^{K} \delta_{j,k} \leq 1, \ j = 1, \ldots, M; \tag{16}$$

$$d_k \geq 0, \delta_{j,k} \geq 0, \ j = 1, \ldots, M, \ k = 1, \ldots, K. \tag{17}$$

Our objective in this LP is to allocate the servers to the classes so that the output from the system is maximized. The right-hand side of the first constraint (14) is the total amount of service effort allocated to class k and the left-hand side is the long-run departure rate from class k. So (14) simply means that the departure rate from a class k cannot exceed the service allocation to that class. Similarly, the right-hand side of constraint (15) is the long-run arrival rate to class k. So this constraint means that the long-run departure rate from a class can not exceed the long-run arrival rate to that class. The constraint (16) prevents us from overallocating a server, and (17) prevents negative allocations.

Let an optimal solution to the above LP for the offered demand λ be given by $\delta^*_{j,k}$ and d^*_k, for all j, k. Let $\mu^*(\lambda) = \sum_{k=1}^{K} d^*_k p_{k,0}$ be the optimal value of the LP corresponding to λ. Clearly, $(d^*_1, \ldots, d^*_K)$ is an optimal solution to the above LP if and only if $(d^*_1, \ldots, d^*_K)$ satisfy the set of equations $(11) - (12)$ with $\delta_{j,k} = \delta^*_{j,k}$, for all j, k. Consequently, one can obtain a solution to $(11) - (12)$ under the optimal allocation $\delta^*_{j,k}$, for all j, k, by solving the LP. The solution to the allocation LP provides an upper bound on the maximum achievable throughput, and we will see that we can get arbitrarily close to this value. The following theorem states this fact; a formal proof will be given in Section 3.3.5. Generalized round robin policies that achieve throughput arbitrarily close to the optimum value of the allocation LP will be described in Sections 3.3.1 and 3.3.2.

Theorem 3.2.1. *(a) Any throughput less than $\mu^*(\lambda)$ can be achieved, where $\mu^*(\lambda)$ is the optimal value of the allocation LP $(13) - (17)$ for the offered demand λ.*

That is, for any given λ and $0 < \epsilon < 1$, there exists a generalized round robin policy π with throughput μ^π such that $\mu^\pi \geq (1-\epsilon)\mu^*(\lambda)$.

(b) A throughput larger than $\mu^*(\lambda)$ cannot be achieved.

We also have a result on the behavior of the optimal objective function value $\mu^*(\lambda)$ as a function of λ.

Lemma 3.2.1. *The optimal objective function value $\mu^*(\lambda)$ obtained from the allocation LP (13)-(17) is a continuous, non-decreasing, piece-wise linear, and concave function of λ.*

Proof. The fact that $\mu^*(\lambda)$ is non-decreasing as we increase λ is obvious, since by increasing λ, we are increasing the feasible set. The concavity and linearity of $\mu^*(\lambda)$ follows from Theorem 5.1 in Bertsimas and Tsitsiklis [12]. Finally, the continuity follows from the concavity. □

3.2.2 Uniqueness

In this section, we show that given the allocations $\delta^*_{j,k},\ j = 1, \ldots, M,\ k = 1, \ldots, K$, the set of equations $(11) - (12)$ has a unique solution (a^*_k, d^*_k) for $k = 1, \ldots, K$. This result has also been proved by Chen and Mandelbaum [23], Lemma 3.2, but we provide a different proof. Note, however, that non-unique allocations may lead to non-unique (a^*_k, d^*_k) values. For instance, consider a network with three classes, external input only to class 1, and with each customer equally likely to go to class 2 or class 3 from class 1, after which they exit the system. We have two servers and three classes with $\mu_{j,k},\ j = 1, 2,\ k = 1, 2, 3$, values given by the (j, k) entry in the matrix

$$H = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 2 & 2 \end{pmatrix}.$$

Let $\lambda = 6$. Then, based on the solution of the allocation LP, $\mu^*(\lambda)$ is 2 and can be achieved through different assignments, each resulting in different (a^*_k, d^*_k) values. For instance, let the $M \times K$ matrix T^* have (j, k) entry $\delta^*_{j,k}$, for all j, k. Consider the following two assignments:

$$T^*_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},\ T^*_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

For both assignments, $\mu^*(\lambda)$ is 2. Then, for the first assignments we have $a^*_2 = a^*_3 = 2.5$ and $d^*_2 = 2$, $d^*_3 = 0$; however for T^*_2, we have $a^*_2 = a^*_3 = 2.5$ and $d^*_2 = 0$, $d^*_3 = 2$.

Consider an arbitrary assignment of servers $\delta_{j,k}$ for all j,k satisfying (16) and (17). For $k=1,\ldots,K$, let the effective processing capacity at class k be $\mu_k=\sum_{j=1}^{M}\mu_{j,k}\delta_{j,k}$, so that (12) yields $d_k=\min(\mu_k,a_k)$. For all $k=1,\ldots,K$, let $w_k=a_k-d_k$ and $z_k=\mu_k-d_k$, so that we have $a_k=w_k+\mu_k-z_k$ and $d_k=\mu_k-z_k$. Substituting a_k and d_k into (11), we obtain

$$w_k+\mu_k-z_k \;=\; \lambda p_{0,k}+\sum_{i=1}^{K}(\mu_i-z_i)p_{i,k}=\lambda p_{0,k}+\sum_{i=1}^{K}\mu_i p_{i,k}-\sum_{i=1}^{K}z_i p_{i,k}. \quad (18)$$

Let $w'=[w_1,\ldots,w_K]$, $z'=[z_1,\ldots,z_K]$, and $q'=[q_1,\ldots,q_K]$, where $'$ denotes the transpose of a matrix, and

$$q_k=\lambda p_{0,k}+\sum_{i=1}^{K}\mu_i p_{i,k}-\mu_k,\ k=1,\ldots,K.$$

Then (18) yields $w-Gz=q$, where $G=I-P'$. If $\mu_k\geq a_k$, then $d_k=a_k$ and $w_k=0$, $z_k\geq 0$. Otherwise, if $\mu_k\leq a_k$, then $z_k=0$, $w_k\geq 0$. Hence in either case we have $w_k z_k=0$, so that we can formulate $(11)-(12)$ as

$$\begin{aligned} w-Gz &= q, && (19)\\ w_k z_k &= 0,\ k=1,\ldots,K, && (20)\\ w_k &\geq 0,\ z_k\geq 0,\ k=1,\ldots,K, && (21)\end{aligned}$$

which is a linear complementarity problem (q,G).

Since $P^n\to 0$ as $n\to\infty$, it follows that G is an M-Matrix (see Chen and Yao [25] Lemma 7.1), and hence G is also a P-Matrix (see Chen and Zhang [26], page 27). Hence, it follows from Theorem 3.15 in Murty [72] (page 213) that (w_k,z_k) in $(19)-(21)$ are uniquely determined for (q,G). Since $\delta_{j,k}$, for all j,k, are fixed, $d_k=\mu_k-z_k$, for all k, are also unique, and so are $a_k=w_k+d_k$, for all k. Hence we have a unique solution for $(11)-(12)$.

3.2.3 Classification of the Nodes

In this section, we identify the stable and unstable sets of nodes based on the solution of the LP. In particular, we separate the nodes into two sets as follows:

$$\begin{aligned} S &= \{k: a_k^*=d_k^*\}, && (22)\\ U &= \{k: a_k^*>d_k^*\}. && (23)\end{aligned}$$

Since there is a unique solution for $(11)-(12)$, see Section 3.2.2, the sets S and U are uniquely determined given the allocations $\{\delta_{j,k}^*\}$. The sets S and U specify the sets

of classes that are stable and unstable, respectively, in the solution of the allocation LP, where a class is defined to be stable if the departure rate from the class equals the arrival rate. Note that the unstable set of classes U cannot simply be determined by comparing the solution of the regular balance equations $\{r_k\}$ with the effective processing rates at each station; i.e., U is in general different from $\{k : r_k \geq \mu_k^*\}$, where $r_k = \lambda p_{0,k} + \sum_{i=1}^{K} r_i p_{i,k}$ and $\mu_k^* = \sum_{j=1}^{M} \mu_{j,k} \delta_{j,k}^*$ for all k.

For example, consider the network illustrated in Figure 1, where all customers arrive to class 1 and each customer is equally likely to either depart or be routed to the other class from each class 1, 2; see also the routing matrix P. Suppose that we have three servers and that the service rates for each class are indicated in the matrix H, where the (j, k) entry is $\mu_{j,k}$:

$$P = \begin{pmatrix} 0 & 0.5 \\ 0.5 & 0 \end{pmatrix}, H = \begin{pmatrix} 6 & 2 \\ 5 & 1 \\ 4 & 0 \end{pmatrix}. \tag{24}$$

Looking at the respective service rates $\mu_{j,k}$ in H, the best assignment of the servers to the classes is not obvious. Since the effective arrival and departure rates at the classes depend on these allocations, identifying the unstable classes from the matrix H by inspection is also not obvious. So we resort to the allocation LP $(13) - (17)$. When $\lambda = 6$, the optimum objective function value $(d_1^*/2 + d_2^*/2)$ is given by $\mu^*(6) \simeq 4.7727$ and the assignments are as follows

$$T^* \simeq \begin{pmatrix} 0 & 1 \\ 0.6364 & 0.3636 \\ 1 & 0 \end{pmatrix}. \tag{25}$$

According to these results, we see that the effective processing capacities, departure, and arrival rates at each class $k = 1, \ldots, K$ are given by

$$\mu^* \simeq [7.1818, 2.3636]', \ d^* \simeq [7.1818, 2.3636]', \ a^* \simeq [7.1818, 3.5909]',$$

where $\mu^* = [\mu_1^*, \ldots, \mu_K^*]'$, $d^* = [d_1^*, \ldots, d_K^*]'$ and $a^* = [a_1^*, \ldots, a_K^*]'$. If we solve the regular balance equations, we obtain $r_1 = 8$ and $r_2 = 4$, so that $\{k : r_k \geq \mu_k^*\} = \{1, 2\}$. However, according to our algorithm, the only unstable class in the solution of the allocation LP is class 2, because we have $a_1^* = d_1^*$ and $a_2^* > d_2^*$, so that $U = \{2\}$ and $S = \{1\}$.

As shown before, we cannot simply determine stable and unstable nodes by inspection, and Jackson networks are no exception. Goodman and Massey [43] identify

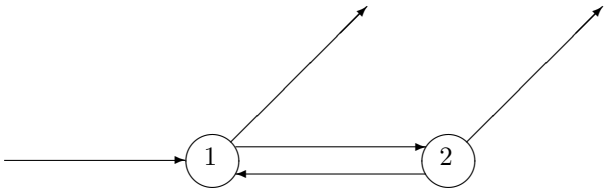


Figure 1: A two-class network

the maximal subnetwork that achieves steady state in a non-ergodic Jackson network. However, the allocation LP can also determine the stable and unstable sets of nodes for Jackson networks with the servers constrained to choose only one class to serve (i.e., $\delta_{j,k} \in \{0, 1\}$, for all j and k). Unlike the algorithm suggested by Goodman and Massey [43] to find the stable and unstable sets, our LP not only provides a classification of the nodes but also suggests an optimal server allocation plan that maximizes throughput. If the server allocation is predetermined (i.e., the $\delta^*_{j,k}$ are given), then the stable and unstable sets suggested by our LP coincide with those determined in Goodman and Massey [43]. Moreover, an invariant distribution exists for the stable set of classes as shown by Goodman and Massey [43].

Note that some policy π with throughput μ^π that comes arbitrarily close to the optimum throughput $\mu^*(\lambda)$ (i.e., $\mu^\pi \geq \mu^*(\lambda) - \epsilon$, where $\epsilon > 0$ is small) does not necessarily have the same sets of stable and unstable classes as determined by the allocation LP. For instance, consider a network with two classes and offered demand $\lambda = 1$, where each job is equally likely to go to class 1 or class 2, from which they exit the system. We have one flexible server with $(\mu_{1,1}, \mu_{1,2}) = (1, 0.5)$. Then, the unique optimal allocations are given by $\delta^*_{1,1} = 1/2$ and $\delta^*_{1,2} = 1/2$ with $\mu^*(\lambda) = 0.75$. Hence the sets S and U are uniquely determined by $\{1\}$ and $\{2\}$, respectively.

Next we consider three allocations that yield different stable and unstable sets. First, for any $0 < \epsilon < 1$, let $(\delta^{(1)}_{1,1}, \delta^{(1)}_{1,2}) = ((1-\epsilon)/2, (1+\epsilon)/2)$, so that $S^{(1)} = \emptyset$, $U^{(1)} = \{1, 2\}$, and $\mu^{(1)} = \mu^*(\lambda) - \epsilon/4 > \mu^*(\lambda) - \epsilon$. Secondly, for any $0 < \epsilon < 1$, let $(\delta^{(2)}_{1,1}, \delta^{(2)}_{1,2}) = ((1+\epsilon)/2, (1-\epsilon)/2)$, so that $S^{(2)} = \{1\}$, $U^{(2)} = \{2\}$, and $\mu^{(2)} = \mu^*(\lambda) - \epsilon/4 > \mu^*(\lambda) - \epsilon$. Finally, consider the assignment $(\delta^{(3)}_{1,1}, \delta^{(3)}_{1,2}) = (0, 1)$, then we have $\mu^{(3)} = 0.5 \geq \mu^*(\lambda) - \epsilon$ for $\epsilon \geq 0.25$, and $S^{(3)} = \{2\}$, $U^{(3)} = \{1\}$. We observe that even if the allocation LP has unique set classifications, we can construct policies based on ϵ with different stable and unstable sets. Hence, we conclude that stability of a class according to the LP does not imply it has to be stable for a near-optimal policy, and vice versa. Returning to the example, if we want to get arbitrarily close

to $\mu^*(\lambda)$ with a small enough ϵ, then only policies 1 and 2 are valid, because the last one violates this requirement. This suggests that as we get closer to the optimum allocations (i.e., $\epsilon \to 0$), the set of unstable classes under a near-optimal policy will contain the set of unstable classes of the allocation LP.

3.3 Optimum Server Allocation

In this section, we develop two alternative server allocation algorithms that achieve throughput that is arbitrarily close to the optimum value of the allocation LP (13) − (17). The analysis is complicated by the observation in the previous section that for a policy π, the set of stable and unstable classes may not correspond to those given by the LP. This makes it difficult to determine the proportion of time spent by a server at each class under π. To ensure that the fraction of time that servers spend at the different classes is sufficiently close to the allocations obtained by the solution of the allocation LP, we propose two approaches. The first, described in Section 3.3.1, involves admission control and controlled routing. The second approach, described in Section 3.3.2, involves forced idling of servers at certain classes. Section 3.3.3 constructs the underlying fluid model for the queueing network described in Section 3.1 and Section 3.3.4 describes a Markov process model for the same queueing network. Section 3.3.5 uses the results from Sections 3.3.3 and 3.3.4 to prove that the algorithms provided in Sections 3.3.1 and 3.3.2 can be used to obtain throughput that is arbitrarily close to the maximum output $\mu^*(\lambda)$ given the available demand λ.

3.3.1 Server Allocation Policy with Admission and Routing Control

In this section, an algorithm for assigning servers to classes is presented based on the allocation LP introduced in Section 3.2.1. In particular, suppose that we are given a certain λ (level of offered demand to the system) and asked to maximize the throughput without regard to stability. Let $\{\delta^*_{j,k}\}$ be the optimal assignment fractions given by the solution to the allocation LP (13) − (17), and let $\mu^*(\lambda) = \sum_{k=1}^{K} d^*_k p_{k,0}$ be the resulting optimum throughput. Our aim is to assign servers to classes based on the fractions $\{\delta^*_{j,k}\}$ to achieve throughput as close to $\mu^*(\lambda)$ as desired. For this, a generalized round robin policy with admission control and controlled routing is considered. More specifically, in this policy, we reject arrivals to the system with a small probability, and also modify the routing probabilities $p_{i,k}$, for all i, k, so that the arrival rate to the classes $k \in U$ is reduced to d^*_k and excess input is rerouted to an imaginary class $K+1$ served by an imaginary server $M+1$. In practice, the customers

routed to class $K+1$ would be scrapped, but the addition of his imaginary class simplifies the analysis as it facilitates differentiation between successful completions and scrapped customers. This policy not only guarantees a target throughput, but also stabilizes the classes in the network by scrapping just enough customers at certain classes in the network.

The following proposition is used to show that for any allocation of servers to classes, a generalized round robin policy exists that gets arbitrarily close to that allocation. For a proof, see Andradóttir, Ayhan, and Down [6], Proposition 3.

Proposition 3.3.1. *Let κ be a finite set, and for each $k \in \kappa$, suppose that m_k and δ_k satisfy $0 < m_k < \infty$, $\delta_k \geq 0$, and $0 \leq \sum_{k\in\kappa} \delta_k \leq 1$. Suppose furthermore that $0 \leq s < \infty$. Then for any $0 < \epsilon \leq 1$, there exists a set of non-negative integers $\{l_k\}$, where $k \in \kappa$, such that*

$$\frac{l_k m_k}{s + \sum_{i\in\kappa} l_i m_i} \geq \delta_k(1-\epsilon) \text{ for all } k \in \kappa. \tag{26}$$

Let $1\{\cdot\}$ denote the indicator function. Then one possible choice for l_k is

$$l_k = \left\lceil \frac{(1-\epsilon)(s + \sum_{i\in\kappa} m_i 1\{\delta_i > 0\})\delta_k}{\epsilon m_k} \right\rceil. \tag{27}$$

Consider a specific policy π that has each server j serving a fixed list V_j^π of classes in a cyclic order. For each class $k \in V_j^\pi$, server j serves a maximum of $l_{j,k}^\pi$ customers and then moves to the next class on the list for service, but if the queue for class k empties before $l_{j,k}^\pi$ service completions, the server moves on to the next class on its list. If there are no more customers in any of the classes on the list, then the server idles until an arrival to any class on the list. We now state how to choose the parameters V_j^π and $l_{j,k}^\pi$ of our generalized round robin server assignment policy π, assuming that the offered demand to the system is λ. In the following algorithm, we are primarily interested in the behavior of the network when $\lambda > \lambda^*$ (i.e., when $U \neq \emptyset$), where λ^* is the maximum offered demand such that the system can be stabilized for $\lambda < \lambda^*$. The case $\lambda < \lambda^*$ is already covered in [6], where it is shown that λ^* can be computed by solving an appropriate LP.

1. Solve the allocation LP (13) − (17).

2. Choose $0 < \epsilon < 1$.

3. Admission Control: Thin the arrival process by rejecting arrivals with probability ϵ and accepting them with probability $1-\epsilon$, so that the arrival rate reduces to $\lambda' = \lambda(1-\epsilon)$.

Controlled Routing: Introduce an imaginary scrapping class $K+1$ with an associated dedicated server $M+1$ such that $\mu_{M+1,K+1} = \lambda$ and $\delta^*_{M+1,K+1} = 1$. Replace the routing probabilities $p_{i,k}$, where $0 \leq i,k \leq K$ by the following routing probabilities $\bar{p}_{i,k}$, $0 \leq i,k \leq K$. For $0 \leq i \leq K$, let $\bar{p}_{i,k} = p_{i,k}$ for $k \in S$; $\bar{p}_{i,k} = p_{i,k}\epsilon_k$ for $k \in U$, where $\epsilon_k = d^*_k/a^*_k$; $\bar{p}_{i,K+1} = \sum_{k\in U} p_{i,k}(1-\epsilon_k)$; and $\bar{p}_{K+1,0} = 1$, $\bar{p}_{K+1,K+1} = 0$. For $1 \leq k \leq K$, $\bar{p}_{k,0} = p_{k,0}$ and $\bar{p}_{K+1,k} = 0$.

4. For each server j, specify the ordered list V^π_j using all of the classes k with $\mu_{j,k}\delta^*_{j,k} > 0$. Define the ith element of each list V^π_j as $v_{j,i}$ and let $|\cdot|$ denote cardinality of a set.

5. For each server j with $|V^\pi_j| > 1$, let s^π_j be the expected switching time in a cycle of visiting the states in V^π_j in order, so that

$$s^\pi_j = \sum_{i=1}^{|V^\pi_j|-1} s^j_{v_{j,i},v_{j,i+1}} + s^j_{v_{j,|V^\pi_j|},v_{j,1}}.$$

6. For each server j with $|V^\pi_j| > 1$ and each class $k \in V^\pi_j$, calculate parameters $l^\pi_{j,k}$ satisfying $l^\pi_{j,k}m_{j,k}/(s^\pi_j + \sum_{i\in V^\pi_j} l^\pi_{j,i}m_{j,i}) \geq \delta^*_{j,k}(1-\epsilon')$, where $\epsilon' = \epsilon/(2-\epsilon)$, see Proposition 3.3.1 and equation (27).

7. For each server j with $|V^\pi_j| = 1$, set $s^\pi_j = 0$ and $l^\pi_{j,k} = 1$ for $k \in V^\pi_j$.

8. For each server j and all classes $k \notin V^\pi_j$, let $l^\pi_{j,k} = 0$.

As a result of ignoring stability in the allocation LP (13) – (17), it is possible to have queue lengths $\{Q_k(t)\}$ at certain classes k diverge as $t \to \infty$, without the controlled routing. The following theorem shows that the above generalized round robin policy π yields throughput μ^π that comes arbitrarily close to achieving the desired throughput level of $\mu^*(\lambda)$, and also stabilizes the original queueing network. The proof of Theorem 3.3.1 is postponed until Section 3.3.5.

Theorem 3.3.1. *A policy constructed using the above algorithm achieves throughput $\mu^\pi = (1-\epsilon)\mu^*(\lambda)$. Moreover, the distribution of the queue length process $\{Q(t)\}$ converges to a steady state distribution as $t \to \infty$.*

It immediately follows from Theorem 3.3.1 that an appropriate value of ϵ will guarantee that we achieve a target throughput $\mu < \mu^*(\lambda)$, as stated in the following corollary.

Corollary 3.3.1. *A policy constructed using the above algorithm with* $\epsilon = 1-\mu/\mu^*(\lambda)$*, where* $\mu < \mu^*(\lambda)$*, achieves a target throughput* μ *(i.e.,* $\mu^\pi = \mu$*).*

3.3.2 Server Allocation Policy with Forced Server Idling

In this section, we introduce an alternative generalized round robin policy without admission control or controlled routing. Since we allow instability in the system, each server j will eventually always find more than the required number of customers $l_{j,k}$ at unstable classes k, and hence spend the maximum amount of time allowed during each of its cycles at such classes in its list. However, this could result in problems, because although the fractions of time servers spend at unstable classes are guaranteed to achieve certain minimums (see Proposition 3.3.1), we do not control how big they can be. Since there are always customers to process at unstable classes, it becomes possible for a server assigned to an unstable class to spend more time than required there, resulting in the flows of customers between stations in the network not being sufficiently close to the optimal flows identified by the allocation LP $(13) - (17)$. To prevent this, we force the servers to spend the required amount of time at each of the classes in their lists, even if it means idling them. Unlike the approach in the previous section where servers complete a fixed number of customers before switching, we will construct a timed round robin policy where servers spend a fixed amount of time at each class on its list. We also assume that service time distributions are independent of the server (this assumption is required since a server may resume a customer service started by another server). Hence, we will represent the service requirement of customer n at class k by $v_k(n)$, and server j reduces this requirement at a rate $\mu_{j,k}$ when assigned to class k.

Consider a specific policy π that has each server j serving a fixed list V_j^π of classes in a cyclic order as in Section 3.3.1. For each class $k \in V_j^\pi$, server j spends a fixed amount of time $h_{j,k}^\pi$ at class k, even if the queue for class k empties before that time, and then server j moves to the next class on its list. We make use of Proposition 3.3.1 to determine $h_{j,k}^\pi$, for all j, k. Although the following algorithm works for any value of λ, we are primarily interested in the behavior of the network when $\lambda > \lambda^*$. Next, we state how to choose the parameters V_j^π and $h_{j,k}^\pi$ of our generalized round robin server assignment policy π, assuming that the offered demand to the system is λ. In particular, we will use the eight-step policy of Section 3.3.1, except that steps 3, 6, and 8 of that policy are replaced by the steps below:

3. For all the servers $j = 1, \ldots, M$, let

$$\kappa_j = 1 - \sum_{k=1}^{K} \delta^*_{j,k} 1\{\mu_{j,k} > 0\}.$$

6. For each server j with $|V_j^\pi| > 1$ and each class $k \in V_j^\pi$, set $\bar{\delta}^*_{j,k} = \delta^*_{j,k} + \kappa_j/|V_j^\pi|$, for all $k \in V_j^\pi$, and calculate parameters $l^\pi_{j,k}$ satisfying $l^\pi_{j,k} m_{j,k}/(s^\pi_j + \sum_{i \in V_j^\pi} l^\pi_{j,i} m_{j,i}) \geq \bar{\delta}^*_{j,k}(1-\epsilon)$, see Proposition 3.3.1 and equation (27).

8. For each server j, set $h^\pi_{j,k} = l^\pi_{j,k} m_{j,k}$, for $k \in V_j^\pi$, and $h^\pi_{j,k} = 0$, for $k \notin V_j^\pi$.

Theorem 3.3.2. *A policy constructed using the above algorithm achieves the throughput $\mu^\pi \geq (1-\epsilon)\mu^*(\lambda)$.*

The proof of Theorem 3.3.2 is provided in Section 3.3.5. It immediately follows from Theorem 3.3.2 that an appropriate value of ϵ will guarantee that we achieve a target throughput $\mu < \mu^*(\lambda)$, as stated in the following corollary.

Corollary 3.3.2. *A policy constructed using the above algorithm with $\epsilon = 1 - \mu/\mu^*(\lambda)$, where $\mu < \mu^*(\lambda)$, achieves a target throughput μ (i.e., $\mu^\pi \geq \mu$).*

3.3.3 A Fluid Model for Queueing Networks

The fluid models involve smoothing out discrete processes, using the SLLN. In this section, we develop a fluid model for the original queueing network described in Section 3.1 under a server assignment policy π. Let $q = \sum_{k=1}^K Q_k(0)$. Suppose that the function $(\bar{Q}_k(\cdot), \bar{T}_{j,k}(\cdot), \forall j,k)$ is a limit point of $(Q_k(qt)/q, T_{j,k}(qt)/q, \forall j,k)$ when $q \to \infty$. Then $(\bar{Q}_k(\cdot), \bar{T}_{j,k}(\cdot) : k = 1, \ldots, K)$ is a fluid limit of the system. Each component of a fluid limit is absolutely continuous (and thus differentiable) almost everywhere in $[0, \infty)$ (see Dai [32], page 20). If we require the derivative of a quantity, we will assume it is taken at a time point t such that the derivative exists (such a point is known as a regular point). For each class $k = 1, \ldots, K$, let $\bar{A}_k(t) = \lim_{q\to\infty} A_k(qt)/q$ and $\bar{D}_k(t) = \lim_{q\to\infty} D_k(qt)/q$ be the fluid limits for the arrival and departure processes $A_k(t)$ and $D_k(t)$, respectively. Then the deterministic analogs $\bar{A}$, $\bar{D}$, and $\bar{Q}$ of the queueing network processes A, D, and Q satisfy the

following equations (see Theorem 4.1 of Dai [33]):

$$\bar{A}_k(t) = \lambda p_{0,k} t + \sum_{i=1}^{K} \sum_{j=1}^{M} p_{i,k} \mu_{j,i} \bar{T}_{j,i}(t), \ k = 1, \ldots, K; \tag{28}$$

$$\bar{D}_k(t) = \sum_{j=1}^{M} \mu_{j,k} \bar{T}_{j,k}(t), \ k = 1, \ldots, K; \tag{29}$$

$$\bar{Q}_k(t) = \bar{Q}_k(0) + \lambda p_{0,k} t + \sum_{i=1}^{K} \sum_{j=1}^{M} p_{i,k} \mu_{j,i} \bar{T}_{j,i}(t) - \sum_{j=1}^{M} \mu_{j,k} \bar{T}_{j,k}(t), \ k = 1, \ldots, K \tag{30}$$

Equations (28) − (30) are obtained from (8) − (10) by replacing $S_{j,k}(t)$, $E_k(t)$, and $\Phi_{i,k}(n)$ by their asymptotic means. Note that (28) − (30) are independent of the selected policy π, the dependence on π is given in the functions $\{\bar{T}_{j,k}(t)\}$.

3.3.4 Underlying Markov Process Construction

In this section, we define a Markov process $X = \{X(t), t > 0\}$ which describes the dynamics of the queueing network described in Section 3.1 with K classes and M servers operating under a generalized round robin policy π, where each server j cycles among all the classes k on its list V_j^π, serving a maximum of $l_{j,k}^\pi$ customers at class k before moving to the next class. Let $U(t)$ and $V_{j,k}(t)$, $j = 1, \ldots, M$, $k = 1, \ldots, K$, be the residual interarrival and service times defined in Section 3.1 and $W_j(t)$ be the residual switching time at time t for server j. Also, let $L_j(t)$ be the location of server j at time t (set to the destination class if the server is switching at time t), $I_j(t)$ be the status of server j (0 if the server is idle or switching, 1 if busy), and $N_j(t)$ be the number of customers finished by server j at the current location $L_j(t)$ (reset to zero each time server j idles or makes a switch). Note that since we have non-preemptive service, the residual service time can be only at the current location $L_j(t)$ at time t, so let $V_j(t)$ be the residual service time for server j. The continuous variables $\{U(t), V_j(t), W_j(t)\}$ are taken to be right continuous. Then the process $X(t)$ defined by

$$X(t) = (U(t), V_j(t), W_j(t), Q_k(t), L_j(t), I_j(t), N_j(t); j = 1, \ldots, M, k = 1, \ldots, K)$$

can be shown to have the strong Markov property as in Section 4 of Davis [34], with elements

$$x \in \mathbb{R}_+ \times \mathbb{R}_+^M \times \mathbb{R}_+^M \times \mathbb{Z}_+^K \times \{1, \ldots, K\}^M \times \{0, 1\}^M \times \{0, 1, \ldots, \max l_{j,k} - 1\}^M.$$

Next, we need to make minor modifications for the allocation policy described in Section 3.3.1 as it results in a slightly modified network. A similar Markov process

exists for the modified network under admission control and controlled routing as for the original network, with the only difference that the dimension of the state is increased by the additional class and server, so that the Markov process evolves on

$$x \in \mathbb{R}_+ \times \mathbb{R}_+^{M+1} \times \mathbb{R}_+^M \times \mathbb{Z}_+^{K+1} \times \{1,\ldots,K\}^M \times \{0,1\}^M \times \{0,1,\ldots,\max_{j,k} l_{j,k} - 1\}^M.$$

Note that $V_{M+1}(t)$ is the only information, we need to keep on the $(M+1)^{th}$ server, because it is a dedicated server assigned to class $K+1$. Since we will not be proving the stability of the queueing network under the policy described in Section 3.3.2, we do not need to describe the resulting Markov process in that case.

3.3.5 Proofs of Theorems 3.2.1, 3.3.1, and 3.3.2

In this section we give formal proofs to Theorems 3.2.1, 3.3.1, and 3.3.2. We start with part (b) of Theorem 3.2.1. Next, we prove Theorem 3.3.1 for the generalized round robin policy introduced in Section 3.3.1. Then part (a) of Theorem 3.2.1 follows. Finally we show that the allocation policy in Section 3.3.2 also achieves the target throughput, as stated in Theorem 3.3.2.

Proof of Theorem 3.2.1(b). We proceed by contradiction. Assume that there exists a policy π and a subset A of the sample space Ω with $P(A) > 0$, such that

$$\limsup_{t\to\infty} \frac{D^\pi(t,\omega)}{t} > \mu^*(\lambda),\ \forall \omega \in A, \tag{31}$$

where $D^\pi(t,\omega)$ is the total number of departures from the system under the policy π in $(0,t]$ for the sample path ω. By the i.i.d. assumption on the primitive processes, there exists a set A' with $P(A') = P(A)$ such that for all $\omega \in A'$, and any $\epsilon, \epsilon_1 > 0$, there exists $T_1(\omega)$ and $N(\omega)$ such that for all $t \geq T_1(\omega)$ and for all $n \geq N(\omega)$, and $i = 0,\ldots,K$, $k = 1,\ldots,K$, and $j = 1,\ldots,M$,

$$\left|\frac{E_k(t,\omega)}{t} - \lambda p_{0,k}\right| \leq \epsilon_1,\ \left|\frac{\Phi_{k,i}(n,\omega)}{n} - p_{k,i}\right| \leq \epsilon_1,\ \left|\frac{S_{j,k}(t,\omega)}{t} - \mu_{j,k}\right| \leq \epsilon.$$

Next we obtain bounds on the cumulative queueing processes, starting with the departure process from each class. We have $D_k(t) = \sum_{j=1}^M S_{j,k}(T_{j,k}(t)),\ k = 1,\ldots,K$. On the sample path $\omega \in A$, some servers may spend a finite amount of time at given classes, resulting in two cases:

- For pairs j,k such that $\lim_{t\to\infty} T_{j,k}(t,\omega) < \infty$, we have $S_{j,k}(T_{j,k}(t),\omega)/t \to 0$, since $S_{j,k}(t,\omega) < \infty$, for all t, by assumption (5).

- For pairs j,k such that $\lim_{t\to\infty} T_{j,k}(t,\omega)=\infty$, we can find $T_2(\omega)$ such that for all $t\geq T_2(\omega)$, we have $T_{j,k}(t,\omega)\geq T_1(\omega)$ implying

$$\left|\frac{S_{j,k}(T_{j,k}(t,\omega))}{T_{j,k}(t,\omega)}-\mu_{j,k}\right|\leq\epsilon.$$

Let $M_k(\omega)=\{j:\mu_{j,k}>0 \text{ and } \lim_{t\to\infty} T_{j,k}(t,\omega)=\infty\}$. We have

$$\frac{D_k(t,\omega)}{t}=\sum_{j\in M_k(\omega)}\frac{S_{j,k}(T_{j,k}(t,\omega))}{T_{j,k}(t,\omega)}\times\frac{T_{j,k}(t,\omega)}{t}+\sum_{j\notin M_k(\omega)}\frac{S_{j,k}(T_{j,k}(t,\omega))}{t},\ k=1,\ldots,K.$$

Let $\delta_{j,k}(t,\omega)=T_{j,k}(t,\omega)/t$. For any $\epsilon_2>0$, there exists $T_3(\omega)$ such that for all $t\geq T_3(\omega)$, we have $\sum_{j\notin M_k(\omega)} S_{j,k}(T_{j,k}(t,\omega))/t\leq\epsilon_2$, for $k=1,\ldots,K$. Then for $t\geq\max\{T_2(\omega),T_3(\omega)\}$, we have

$$\frac{D_k(t,\omega)}{t}\leq\sum_{j\in M_k(\omega)}(\mu_{j,k}+\epsilon)\delta_{j,k}(t,\omega)+\epsilon_2,\ k=1,\ldots,K.$$

Let $\epsilon_3=\epsilon M+\epsilon_2$, which implies that $\epsilon_3\geq\max_k\{\epsilon\sum_{j\in M_k(\omega)}\delta_{j,k}(t,\omega)+\epsilon_2\}$, and thus for $t\geq\max\{T_2(\omega),T_3(\omega)\}$, we obtain

$$\frac{D_k(t,\omega)}{t}-\epsilon_3\leq\sum_{j\in M_k(\omega)}\mu_{j,k}\delta_{j,k}(t,\omega),\ k=1,\ldots,K. \tag{32}$$

Next, we bound the arrival process to each class. Let $\mathcal{K}=\{1,\ldots,K\}$ denote the set of all classes in the network, and define

$$\mathcal{K}\setminus\bar{\mathcal{K}}(\omega)=\{k:\lim_{t\to\infty}T_{j,k}(t,\omega)<\infty,\forall j \text{ with } \mu_{j,k}>0\}=\{k:M_k(\omega)=\emptyset\}.$$

Note that all of the servers capable of working at the classes in $\mathcal{K}\setminus\bar{\mathcal{K}}(\omega)$ spend only a finite amount of time at those classes, so that the number of departures is bounded. For the arrival process, we have

$$A_k(t,\omega)=E_k(t,\omega)+\sum_{i=1}^{K}\Phi_{i,k}(D_i(t,\omega)),\ k=1,\ldots,K.$$

For $i\in\bar{\mathcal{K}}(\omega)$, $\lim_{t\to\infty} D_i(t,\omega)=\infty$, and hence there exists $T_4(\omega)$ such that for all $t\geq T_4(\omega)$, we have $D_i(t,\omega)>N(\omega)$, implying

$$\left|\frac{\Phi_{i,k}(D_i(t,\omega))}{D_i(t,\omega)}-p_{i,k}\right|\leq\epsilon_1,\ k=0,1,\ldots,K.$$

For $i\in\mathcal{K}\setminus\bar{\mathcal{K}}(\omega)$, $\lim_{t\to\infty} D_i(t,\omega)<\infty$, and $\lim_{t\to\infty}\Phi_{i,k}(D_i(t,\omega))/t=0$. Hence, for any $\epsilon_4>0$, there exists $T_5(\omega)$ such that for all $t\geq T_5(\omega)$, we have

$$\sum_{i\in\mathcal{K}\setminus\bar{\mathcal{K}}(\omega)}\frac{\Phi_{i,k}(D_i(t,\omega))}{t}\leq\epsilon_4,\ k=0,1,\ldots,K.$$

Then, for the arrival process, we have for $t \geq \max\{T_1(\omega), T_4(\omega), T_5(\omega)\}$,

$$\frac{A_k(t,\omega)}{t} \leq \lambda p_{0,k} + \epsilon_1 + \sum_{i \in \bar{\mathcal{K}}(\omega)} (p_{i,k} + \epsilon_1) \frac{D_i(t,\omega)}{t} + \epsilon_4, \ k = 1, \ldots, K.$$

Plugging in (32), we get, for $t \geq \max\{T_1(\omega), T_2(\omega), T_3(\omega), T_4(\omega), T_5(\omega)\}$,

$$\frac{A_k(t,\omega)}{t} \leq \lambda p_{0,k} + \sum_{i \in \bar{\mathcal{K}}(\omega)} p_{i,k} \frac{D_i(t,\omega)}{t} + \epsilon_1 \sum_{i \in \bar{\mathcal{K}}(\omega)} \sum_{j \in M_i(\omega)} \delta_{j,i}(t,\omega) \mu_{j,i} + \epsilon_1 \sum_{i \in \bar{\mathcal{K}}(\omega)} \epsilon_3 + \epsilon_4 + \epsilon_1, \ k = 1, \ldots, K.$$

We also have $D_k(t,\omega) \leq A_k(t,\omega) + Q_k(0)$ for all $t \geq 0$. Let $\epsilon_5 = \epsilon_1 K M \mu + K \epsilon_1 \epsilon_3 + \epsilon_4 + 2\epsilon_1$, where $\mu = \max\{\mu_{j,i}, j = 1, \ldots, M, i = 1, \ldots, K\}$, so that

$$\epsilon_5 \geq \max_k \Big\{ \epsilon_1 \sum_{i \in \bar{\mathcal{K}}(\omega)} \sum_{j \in M_i(\omega)} \delta_{j,i}(t,\omega) \mu_{j,i} + \epsilon_1 \sum_{i \in \bar{\mathcal{K}}(\omega)} \epsilon_3 + \epsilon_4 + 2\epsilon_1 \Big\}.$$

Then for $t \geq \max\{T_1(\omega), T_2(\omega), T_3(\omega), T_4(\omega), T_5(\omega), Q_k(0)/\epsilon_1\}$ we have

$$\frac{D_k(t,\omega)}{t} - \epsilon_5 \leq \lambda p_{0,k} + \sum_{i \in \bar{\mathcal{K}}(\omega)} p_{i,k} \frac{D_i(t,\omega)}{t}, \ k = 1, \ldots, K. \tag{33}$$

Finally, we bound the departure process from the system, $D(t,\omega) = \sum_{i=1}^{K} \Phi_{i,0}(D_i(t,\omega))$. For $t \geq \max\{T_4(\omega), T_5(\omega)\}$, we have

$$\frac{D(t,\omega)}{t} \leq \sum_{i \in \bar{\mathcal{K}}(\omega)} (p_{i,0} + \epsilon_1) \times \frac{D_i(t,\omega)}{t} + \epsilon_4.$$

Let $\epsilon_6 = \epsilon_1 K(\epsilon_3 + M\mu) + \epsilon_4$, so that (32) implies that $\epsilon_6 \geq \epsilon_1 \sum_{i \in \bar{\mathcal{K}}(\omega)} D_i(t,\omega)/t + \epsilon_4$. Then we get for $t \geq \max\{T_4(\omega), T_5(\omega)\}$

$$\frac{D(t,\omega)}{t} - \epsilon_6 \leq \sum_{i \in \bar{\mathcal{K}}(\omega)} p_{i,0} \frac{D_i(t,\omega)}{t}. \tag{34}$$

By assumption, under policy π, the departure process satisfies (31). Let $l = \limsup_{t \to \infty} D^\pi(t,\omega)/t > \mu^*(\lambda)$. Then for any $\epsilon_7 > 0$, $D^\pi(t,\omega)/t \geq l - \epsilon_7$ infinitely often. Then we can choose a time $t_0 \geq \max\{T_1(\omega), T_2(\omega), T_3(\omega), T_4(\omega), T_5(\omega), Q_k(0)/\epsilon_1\}$ with an ϵ_7 small enough so that $D^\pi(t_0,\omega)/t_0 > \mu^*(\lambda)$ and also the bounds in (32), (33) and (34) are satisfied at t_0. Rewriting (32) – (34) for the cumulative processes

at time t_0, we get

$$\frac{D_k(t_0,\omega)}{t_0} - \epsilon_3 \leq \sum_{j \in M_k(\omega)} \mu_{j,k}\delta_{j,k}(t_0,\omega),\ k = 1,\ldots,K, \tag{35}$$

$$\frac{D_k(t_0,\omega)}{t_0} - \epsilon_5 \leq \lambda p_{0,k} + \sum_{i \in \bar{\mathcal{K}}(\omega)} p_{i,k}\frac{D_i(t_0,\omega)}{t_0},\ k = 1,\ldots,K, \tag{36}$$

$$\frac{D(t_0,\omega)}{t_0} - \epsilon_6 \leq \sum_{i \in \bar{\mathcal{K}}(\omega)} p_{i,0}\frac{D_i(t_0,\omega)}{t_0}. \tag{37}$$

Next, our aim is to show that given the above bounds on the cumulative processes, there exists a solution to the LP $(13)-(17)$ with an objective value greater than $\mu^*(\lambda)$, which will yield the desired contradiction. To see this, define

$$d_k = \begin{cases} 0 & \text{if } k \in K \setminus \bar{\mathcal{K}}(\omega), \\ \frac{D_k(t_0,\omega)}{t_0} & \text{if } k \in \bar{\mathcal{K}}(\omega), \end{cases}$$

and

$$\delta_{j,k} = \begin{cases} 0 & \text{if } j \notin M_k(\omega), \\ \delta_{j,k}(t_0,\omega) & \text{if } j \in M_k(\omega). \end{cases}$$

Plugging these in (35), (36) and (37) and noting that the bounds in $(35)-(37)$ hold for arbitrarily small ϵ_3, ϵ_5, and ϵ_6, respectively, we get after a little manipulation

$$\sum_{i=1}^{K} p_{i,0}d_i > \mu^*(\lambda), \tag{38}$$

$$d_k \leq \sum_{j=1}^{M} \mu_{j,k}\delta_{j,k},\ k = 1,\ldots,K, \tag{39}$$

$$d_k \leq \lambda p_{0,k} + \sum_{i=1}^{K} p_{i,k}d_i,\ k = 1,\ldots,K. \tag{40}$$

By definition, we have $d_k \geq 0$ and $\delta_{j,k} \geq 0$. Moreover, $\sum_{k=1}^{K} T_{j,k}(t_0,\omega) \leq t_0$ implies that $\sum_{k=1}^{K} \delta_{j,k} \leq 1$ for $j = 1,\ldots,M$. Then we see that along this sample path ω under the policy π, we can construct a solution to the LP $(13)-(17)$ with an objective value greater than $\mu^*(\lambda)$, a contradiction. □

Proof of Theorem 3.3.1 and Theorem 3.2.1(a). We will refer to the network obtained as a result of the controlled routing in step 3 of the policy π described in Section 3.3.1 as the "modified" queueing network. Hence step 3 of this policy results in a modified network under admission control. Let $\bar{P}$ be the routing matrix for the modified network (so that $\bar{P}$ has (i,k) entry $\bar{p}_{i,k}$ for $i,k = 1,\ldots,K+1$). Then we have that

$(I - \bar{P})$ is invertible and $\bar{P}^n \to 0$ since the modified network is open (see, e.g., Lawler [64], page 27).

To prove Theorem 3.3.1, we need to develop the queueing network equations and the corresponding fluid model for the modified network. We can obtain these in the same way as we did in Sections 3.1 and 3.3.3. So we have

$$\begin{aligned} A_k(t) &= E_k(t) + \sum_{i=1}^{K+1} \bar{\Phi}_{i,k}(D_i(t)),\ k = 1, \ldots, K+1; \\ D_k(t) &= \sum_{j=1}^{M+1} S_{j,k}(T_{j,k}(t)),\ k = 1, \ldots, K+1; \\ Q_k(t) &= Q_k(0) + A_k(t) - D_k(t),\ k = 1, \ldots, K+1; \end{aligned}$$

and $0 \leq \sum_{k=1}^{K+1} T_{j,k}(t) \leq t,\ j = 1, \ldots, M+1$, where $\bar{\Phi}_{i,k}(n) = \sum_{l=1}^{n} \bar{\phi}_{i,k}(l)$ and the random variables $\bar{\phi}_{i,k}(l)$ are independent and have value one with probability $\bar{p}_{i,k}$ and are zero otherwise. Similarly, fluid limits $\bar{A}_k(t)$, $\bar{D}_k(t)$, and $\bar{Q}_k(t)$ for the modified network under admission control are defined in the same manner as for the original network, for $k = 1, \ldots, K+1$, and satisfy the equations

$$\begin{aligned} \bar{A}_k(t) &= \lambda' \bar{p}_{0,k} t + \sum_{i=1}^{K+1} \sum_{j=1}^{M+1} \bar{p}_{i,k} \mu_{j,i} \bar{T}_{j,i}(t),\ k = 1, \ldots, K+1; \\ \bar{D}_k(t) &= \sum_{j=1}^{M+1} \mu_{j,k} \bar{T}_{j,k}(t),\ k = 1, \ldots, K+1; \\ \bar{Q}_k(t) &= \bar{Q}_k(0) + \lambda' \bar{p}_{0,k} t + \sum_{i=1}^{K+1} \sum_{j=1}^{M+1} \bar{p}_{i,k} \mu_{j,i} \bar{T}_{j,i}(t) - \sum_{j=1}^{M+1} \mu_{j,k} \bar{T}_{j,k}(t), \\ &\qquad\qquad k = 1, \ldots, K+1; \end{aligned} \tag{41}$$

subject to the conditions

$$0 \leq \sum_{k=1}^{K+1} \bar{T}_{j,k}(t) \leq t,\ j = 1, \ldots, M+1;$$

$\bar{T}_{j,k}(0) = 0$ and $\bar{T}_{j,k}(\cdot)$ is non-decreasing for $j = 1, \ldots, M+1, k = 1, \ldots, K+1$;

$$\bar{Q}_k(t) \geq 0, \qquad k = 1, \ldots, K+1;$$

$$\frac{l^{\pi}_{j,k} m_{j,k}}{s^{\pi}_j + \sum_{i \in V^{\pi}_j} l^{\pi}_{j,i} m_{j,i}} \leq \frac{d\bar{T}_{j,k}(t)}{dt} \leq 1, \qquad j = 1, \ldots, M+1,$$

$$k = 1, \ldots, K+1, \text{ whenever } \bar{Q}_k(t) > 0. \tag{42}$$

The lower bound in (42) can be derived as in Andradóttir, Ayhan, and Down [6].

Let $r_k = d_k^*$, for $k = 1, \ldots, K$, and $r_{K+1} = \lambda \bar{p}_{0,K+1} + \sum_{i=1}^{K} d_i^* \bar{p}_{i,K+1}$. We claim that $r_1, \ldots, r_{K+1}$ satisfy the traffic equations for the modified queueing network under the offered demand λ, so that

$$r_k = \lambda \bar{p}_{0,k} + \sum_{i=1}^{K+1} r_i \bar{p}_{i,k}, \ k = 1, \ldots, K+1. \tag{43}$$

To see this, recall that $a_k^* = \lambda p_{0,k} + \sum_{i=1}^{K} d_i^* p_{i,k}$, for $k = 1, \ldots, K$. First consider the classes $k \in U$. Then, we get

$$\begin{aligned} \lambda \bar{p}_{0,k} + \sum_{i=1}^{K+1} r_i \bar{p}_{i,k} &= \lambda p_{0,k} \epsilon_k + \sum_{i=1}^{K} d_i^* p_{i,k} \epsilon_k \\ &= \epsilon_k (\lambda p_{0,k} + \sum_{i=1}^{K} d_i^* p_{i,k}) = \frac{d_k^*}{a_k^*} a_k^* = d_k^* = r_k, \ k \in U, \end{aligned}$$

as required. Next consider the classes $k \in S$. Then $d_k^* = a_k^*$, and we get

$$\lambda \bar{p}_{0,k} + \sum_{i=1}^{K+1} r_i \bar{p}_{i,k} = \lambda p_{0,k} + \sum_{i=1}^{K} d_i^* p_{i,k} = a_k^* = d_k^* = r_k, \ k \in S.$$

Finally, $r_{K+1} = \lambda \bar{p}_{0,K+1} + \sum_{i=1}^{K+1} r_i \bar{p}_{i,K+1}$ follows from the definition of r_{K+1} and the fact that $\bar{p}_{K+1,K+1} = 0$. We have shown that $r_1, \ldots, r_{K+1}$ satisfy (43), and since $I - \bar{P}$ is invertible, the solution is unique.

Let α_k, $k = 1, \ldots, K+1$, be the unique solution of the system of equations (43) when $\lambda = 1$. Then we have $\alpha_k = d_k^*/\lambda$, $k = 1, \ldots, K$, and $\alpha_{K+1} = \bar{p}_{0,K+1} + (\sum_{i=1}^{K} d_i^* \bar{p}_{i,K+1})/\lambda$. Moreover, by constraint (14) in the allocation LP, we have

$$r_k = d_k^* \ \leq \ \sum_{j=1}^{M} \mu_{j,k} \delta_{j,k}^* = \sum_{j=1}^{M+1} \mu_{j,k} \delta_{j,k}^*, \ k = 1, \ldots, K, \tag{44}$$

and

$$r_{K+1} \ \leq \ \lambda = \sum_{j=1}^{M+1} \mu_{j,K+1} \delta_{j,K+1}^* \tag{45}$$

follows from the facts that $p_{K+1,0} = 1$, r_{K+1} is the flow through node $K+1$ in a stable queueing network with offered demand λ and routing matrix $\bar{P}$, and if $r_{K+1} > \lambda$, the system would have more output than input. Then, by the server allocation policy π, and (44) – (45), we have, for all $k = 1, \ldots, K+1$,

$$\sum_{j=1}^{M+1} \frac{l_{j,k}^{\pi}}{s_j^{\pi} + \sum_{i \in V_j^{\pi}} l_{j,i}^{\pi} m_{j,i}} \geq \sum_{j=1}^{M+1} \mu_{j,k} \delta_{j,k}^* (1 - \epsilon') \geq r_k (1 - \epsilon'). \tag{46}$$

Let $\lambda' = \lambda(1-\epsilon)$ be the thinned offered demand, and $r'_k = \lambda'\alpha_k$, $k = 1, \ldots, K+1$, be the solution to the traffic equations for the modified network corresponding to the offered demand λ'. Since $(1-\epsilon')/(1-\epsilon) = 1+\epsilon'$, we have

$$r_k(1-\epsilon') = \lambda\alpha_k(1-\epsilon') = \lambda'\alpha_k\frac{1-\epsilon'}{1-\epsilon} = r'_k(1+\epsilon'),\ k = 1, \ldots, K+1. \tag{47}$$

Plugging (47) in (46), we get

$$\sum_{j=1}^{M+1}\frac{l^\pi_{j,k}}{s^\pi_j + \sum_{i\in V^\pi_j} l^\pi_{j,i}m_{j,i}} \geq r'_k(1+\epsilon'),\ k = 1, \ldots, K+1. \tag{48}$$

Equations (42) and (48) for the modified network imply that when $\bar{Q}_k(t) > 0$, $\sum_{j=1}^{M+1}\frac{d\bar{T}_{j,k}(t)}{dt} \geq r'_k(1+\epsilon')$. By Theorem 2.4.9 of Dai [32], this means that there is a finite time t_0 such that the system is empty and the fluid model for the modified network is stable under the offered demand λ'. Then by Theorem 4.2 of Dai [33], the Markov chain describing the dynamics of the modified network is positive Harris recurrent. Hence, the modified queueing network is stable for the offered demand λ', and the distribution of the queue length process $\{Q_k(t)\}$, $k = 1, \ldots, K+1$, converges to a steady state limit as $t \to \infty$.

Finally, it remains to find the throughput μ^π for the modified network with offered demand λ' under the policy π, i.e., without the customers serviced at class $K+1$. For this, consider the fluid scale queue length differential equation obtained from (41) for the modified network under admission control. Given the queueing network is stable, there exists some time t_0 such that $\sum_{k=1}^{K+1}\bar{Q}_k(t) = 0$ for $t \geq t_0$. Then, for any $t > t_0$, we have

$$0 = \lambda'\bar{p}_{0,k} + \sum_{i=1}^{K+1}\bar{p}_{i,k}\sum_{j=1}^{M+1}\mu_{j,i}\frac{d\bar{T}_{j,i}(t)}{dt} - \sum_{j=1}^{M+1}\mu_{j,k}\frac{d\bar{T}_{j,k}(t)}{dt},\ k = 1, \ldots, K+1. \tag{49}$$

Let $\bar{d}_k(t) = d\bar{D}_k(t)/dt = \sum_{j=1}^{M+1}\mu_{j,k}d\bar{T}_{j,k}(t)/dt$ be the fluid level departure rate from class k, for $k = 1, \ldots, K+1$, in the above equation (49). Then we see that solving the set of equations (49) for $\bar{d}_k(t)$, $k = 1, \ldots, K+1$, gives the same solution as for the traffic equations in (43) when the offered demand is λ'. Hence, $\bar{d}_k(t)$, $k = 1, \ldots, K+1$, are uniquely given by $\bar{d}_k(t) = r'_k$, $k = 1, \ldots, K+1$, and the fluid level total throughput rate $\bar{d}(t) = d\bar{D}(t)/dt$ from classes $k = 1, \ldots, K$ is

$$\bar{d}(t) \;=\; \sum_{k=1}^{K}\bar{p}_{k,0}\sum_{j=1}^{M+1}\mu_{j,k}\frac{d\bar{T}_{j,k}(t)}{dt} = \sum_{k=1}^{K}r'_k\bar{p}_{k,0} = \sum_{k=1}^{K}(1-\epsilon)d^*_k p_{k,0} = \mu^*(\lambda)(1-\epsilon),$$

and hence

$$\bar{D}(t) - \bar{D}(t_0) = \mu^*(\lambda)(1-\epsilon)(t-t_0). \tag{50}$$

Connecting back to the queueing network, recall that $\bar{D}(t)$ is a limit point of $D^\pi(qt)/q$ as $q \to \infty$, where $D^\pi(t) = \sum_{k=1}^K \bar{\Phi}_{k,0}(D_k(t))$ is the total number of departures from the modified network until time t from classes $k = 1, \ldots, K$ with offered demand λ' under the policy π. Assume $l = \limsup_{t\to\infty} D^\pi(t)/t \neq \mu^*(\lambda)(1-\epsilon)$. Then, there exists a sequence $\{t_k\}$ such that $\lim_{k\to\infty} D^\pi(t_k)/t_k = l$. Hence, there exists a fluid limit $\bar{D}(\cdot)$ such that $\bar{D}(t) = \lim_{q\to\infty} tD^\pi(qt)/qt = tl$, contradicting (50). So we have

$$\limsup_{t\to\infty} \frac{D^\pi(t)}{t} = \mu^*(\lambda)(1-\epsilon).$$

This completes the proof for Theorem 3.3.1, and part (a) of Theorem 3.2.1 immediately follows. □

Proof of Theorem 3.3.2. By Steps 3 and 6 of the generalized round robin policy π in Section 3.3.2, we obtain an alternative feasible solution to the LP (13) − (17), that is also optimal. By inflating some $\delta^*_{j,k}$, we are relaxing some of the bounds in (14) and (15) and also making each of the constraints in (16) tight. Let $\bar{d}^*_k$ be the corresponding departure rates with allocation $\bar{\delta}^*_{j,k}$, see (12). Then we see that $\bar{d}^*_k \geq d^*_k$, hence this feasible solution also achieves the optimal. From now on, we will refer to the alternative LP solution $\bar{d}^*_k$, $\bar{\delta}^*_{j,k}$ as d^*_k, $\delta^*_{j,k}$.

As a result of the policy π, each server spends exactly the same amount of time at any class during each cycle of visiting the classes in its list. Let $I_{j,k}(t)$ be the cumulative idle time for server j at class k, and $\bar{I}_{j,k}(t)$ the corresponding fluid limit. Then the fluid model for the queueing network under the server allocation policy of Section 3.3.2 satisfies the equations (28) − (30) subject to the conditions

$$0 \leq \sum_{k=1}^K \bar{T}_{j,k}(t) \leq t, \ j = 1, \ldots, M;$$

$$\sum_{k=1}^K \bar{T}_{j,k}(t) + \bar{I}_{j,k}(t) = t, \ j = 1, \ldots, M;$$

$\bar{T}_{j,k}(0) = 0$, $\bar{I}_{j,k}(0) = 0$, and $\bar{T}_{j,k}(\cdot)$ and $\bar{I}_{j,k}(\cdot)$ are non-decreasing for $j = 1, \ldots, M$, $k = 1, \ldots, K$;

$$\bar{Q}_k(t) \geq 0, \text{ and } \bar{Q}_k(t)\frac{d\bar{I}_{j,k}(t)}{dt} = 0, \ k = 1, \ldots, K; \tag{51}$$

$$\frac{d\bar{T}_{j,k}(t)}{dt} + \frac{d\bar{I}_{j,k}(t)}{dt} = \frac{h^{\pi}_{j,k}}{s^{\pi}_j + \sum_{i \in V^{\pi}_j} h^{\pi}_{j,i}},\ j = 1, \ldots, M,\ k = 1, \ldots, K; \text{ and}$$

$$\frac{d\bar{T}_{j,k}(t)}{dt} = \frac{h^{\pi}_{j,k}}{s^{\pi}_j + \sum_{i \in V^{\pi}_j} h^{\pi}_{j,i}},\ j = 1, \ldots, M,\ k = 1, \ldots, K, \text{ whenever } \bar{Q}_k(t) > 0.$$

The second constraint in equation (51) means that $\bar{I}_{j,k}(t)$ can only increase when $\bar{Q}_k(t)$ is zero. Whenever the amount of fluid at a given class k is positive, then the fluid level is decreased at a constant rate by each server j such that $h^{\pi}_{j,k} > 0$. Then $\sum_{j:k \in V^{\pi}_j} (h^{\pi}_{j,k}\mu_{j,k})/(s^{\pi}_j + \sum_{i \in V^{\pi}_j} h^{\pi}_{j,i})$ is the total rate at which the fluid level at class k is decreased whenever $\bar{Q}_k(t) > 0$ for $k = 1, \ldots, K$.

Next consider a fixed server system with K servers operating under the non-idling FCFS service discipline, external arrival rate λ, and routing probabilities among the classes given in the matrix P. Assume that server k is assigned to class k, and set the service rates μ^{π}_k, $k = 1, \ldots, K$, of the servers as

$$\mu^{\pi}_k = \sum_{j:k \in V^{\pi}_j} \frac{h^{\pi}_{j,k}\mu_{j,k}}{s^{\pi}_j + \sum_{i \in V^{\pi}_j} h^{\pi}_{j,i}}.$$

Then we see that this fixed server system has fluid limits $\{\bar{A}_k(t), \bar{D}_k(t), \bar{Q}_k(t), \forall k \in K\}$, satisfying the same properties as the multi-class system with M servers operating under the policy of Section 3.3.2. Since we have the same routing matrix P for both systems, this means the fluid limits for the throughput $\bar{D}(t) = \sum_{k=1}^{K} \bar{D}_k(t)p_{k,0}$, and hence the throughput of the original system, are equal.

To analyze the throughput in the fixed server system, we can proceed as in Chen and Mandelbaum [23]. Let a_k and d_k be the arrival and departure rates (defined as the inflow and outflow capacities in [23]) at servers $k = 1, \ldots, K$ with the corresponding vectors A and D. Let μ be the K-dimensional processing capacity at each server, with the k^{th} element μ^{π}_k. Then A, D, μ, and the external arrival rate vector E with $E_k = \lambda p_{0,k}$, $k = 1, \ldots, K$, satisfy the traffic equations,

$$A = E + P'D, \tag{52}$$

$$D = A \wedge \mu, \tag{53}$$

where $\wedge$ denotes the componentwise minimum. We know from Section 3.2.2 that (52) − (53) has a unique solution for A and D, when μ is given. The throughput of the fixed server system $\tau(\mu)$ as a function of the processing capacity μ is given by

$$\tau(\mu) = \sum_{i=1}^{K} d_i p_{i,0} = e'(I - P')(A \wedge \mu),$$

where e is the K-dimensional unit vector, see page 426 of Chen and Mandelbaum [23].

Now, $\tau(\mu)$ is a nondecreasing function of the processing capacity μ (see page 427 of Chen and Mandelbaum [23]). Let μ^* be the vector of processing capacities corresponding to the optimal allocations, so that the k^{th} entry of μ^* is $\mu_k^* = \sum_{j=1}^{M} \mu_{j,k}\delta_{j,k}^*$, $k = 1, \ldots, K$. Then we see that (52)−(53) are satisfied for $a_k = a_k^*$ and $d_k = d_k^*$. Hence the maximum throughput for the fixed server system with processing capacity μ^* is given by $\tau(\mu^*) = \sum_{i=1}^{K} d_i^* p_{i,0} = \mu^*(\lambda)$. By Proposition 3.3.1, we have $\mu \geq \mu^*(1-\epsilon)$ so that $\tau(\mu) \geq \tau(\mu^*(1-\epsilon))$. We claim that for the fixed server system with processing capacity $\mu^*(1-\epsilon)$, the throughput $\tau(\mu^*(1-\epsilon))$ is at least $\mu^*(\lambda)(1-\epsilon)$. This follows because we have

$$d_k^*(1-\epsilon) \leq \mu_k^*(1-\epsilon), \tag{54}$$

$$d_k^*(1-\epsilon) \leq \lambda(1-\epsilon)p_{0,k} + (1-\epsilon)\sum_{i=1}^{K} d_i^* p_{i,k} \leq \lambda p_{0,k} + \sum_{i=1}^{K} d_i^*(1-\epsilon)p_{i,k}. \tag{55}$$

Inequality (54) follows from (14) and (55) follows from (15). But then $d_k' = d_k^*(1-\epsilon)$ is a feasible solution for the allocation LP (13)−(17) with fixed servers having processing capacity $\mu^*(1-\epsilon)$, and d_k is the optimal solution. Hence, we have $\tau(\mu) \geq \tau(\mu^*(1-\epsilon)) = \sum_{k=1}^{K} d_k p_{k,0} \geq \sum_{k=1}^{K} d_k' p_{k,0} = \mu^*(\lambda)(1-\epsilon)$ as required. □

3.4 The Saturation Input and Maximum Output

Even if we allow some of the classes in the network to be unstable, the output from the network does not necessarily increase with increased offered demand λ. We refer to the point $\bar{\lambda}$ where increasing the offered demand has no effect on the best possible output as the "saturation" input to the system, and we let $\bar{\mu}$ denote the corresponding maximum output. In this section, we discuss how to identify $\bar{\lambda}$ and $\bar{\mu}$. This information determines the limitations for our system. We also show how to determine the minimum demand required based on a target output level of $\mu \leq \bar{\mu}$.

To determine $\bar{\mu}$, we use the allocation LP $(13) - (17)$ with the only difference that

we set $\lambda = \infty$:

$$\max \sum_{k=1}^{K} d_k p_{k,0} \text{ such that}$$

$$\begin{aligned} d_k &\leq \sum_{j=1}^{M} \mu_{j,k}\delta_{j,k},\ k = 1, \ldots, K; && (56)\\ d_k &\leq \sum_{i=1}^{K} d_i p_{i,k},\ \forall k : p_{0,k} = 0; && (57)\\ \sum_{k=1}^{K} \delta_{j,k} &\leq 1,\ j = 1, \ldots, M;\\ d_k &\geq 0,\ \delta_{j,k} \geq 0,\ j = 1, \ldots, M,\ k = 1, \ldots, K. && (58) \end{aligned}$$

The following theorem shows that the solution of this LP allows us to identify the maximum output $\bar{\mu}$ and to define an upper bound on the saturation input $\bar{\lambda}$.

Theorem 3.4.1. *(a) Let $\bar{\mu} = \sum_{k=1}^{K} d_k^* p_{k,0}$ be the optimal value for the allocation LP (56) − (58) and*

$$\hat{\lambda} = \max_{k:p_{0,k}>0} \left\{ \frac{d_k^* - \sum_{i=1}^{K} d_i^* p_{i,k}}{p_{0,k}} \right\}. \qquad (59)$$

Then we have $\bar{\lambda} \leq \hat{\lambda}$ and $\mu^*(\lambda) = \bar{\mu}$, for all $\lambda \geq \hat{\lambda}$. That is, even if the arrival rate to the original queueing network is increased beyond $\hat{\lambda}$, any capacity larger than $\bar{\mu}$ can not be achieved.

(b) The optimal value $\bar{\mu}$ of the allocation LP (56) − (58) is a tight upper bound on the maximum achievable throughput. That is, any capacity larger than $\bar{\mu}$ cannot be achieved in the original queueing network. Moreover, given a demand $\lambda \geq \hat{\lambda}$, there exists a specific round robin policy π with parameters given by the solution of the LP (56) − (58) and constructed as in Section 3.3.1 or Section 3.3.2 with $\mu^\pi \geq \bar{\mu}(1 - \epsilon)$, where $0 < \epsilon < 1$.

Proof. The optimum value $\bar{\mu}$ of the allocation LP (56) − (58) is finite, since (56) implies that $d_k \leq \sum_{j=1}^{M} \mu_{j,k}$, $k = 1, \ldots, K$, and $\sum_{k=1}^{K} d_k p_{k,0} \leq \sum_{k=1}^{K} d_k$. Also note that $\delta_{j,k}^*$, d_k^* from the solution of the above LP also satisfy the allocation LP (13)−(17) for any $\lambda \geq \hat{\lambda}$ with an optimum value $\mu^*(\lambda) = \bar{\mu}$, since (15) is automatically satisfied by definition of $\hat{\lambda}$. Together with part (b) of Theorem 3.2.1, this proves part (a) of the theorem.

By Theorem 3.2.1, we know that $\bar{\mu}$ is a tight upper bound on the achievable throughput. Moreover, Theorems 3.3.1 and 3.3.2 show that a policy π constructed as in Section 3.3.1 or 3.3.2 will achieve $\mu^\pi \geq \bar{\mu}(1-\epsilon)$, and part (b) of the theorem follows. □

Next our aim is to show how to determine a policy based on a target throughput and also to show how to find the saturation input $\bar{\lambda}$. Because of the non-uniqueness of optimal solutions, $\hat{\lambda}$ can be different from $\bar{\lambda}$. For instance, consider a network with two stations in tandem, each having exactly one dedicated server with processing rates μ_1 and μ_2, respectively. Suppose furthermore that $\lambda > \mu_1 > \mu_2$. Then $d_1^* = \mu_1$, $d_2^* = \mu_2$ is an optimal solution with $\hat{\lambda} = \mu_1$, but $\bar{\lambda} = \mu_2$. We need this tighter saturation input bound to gain insight into the limitations of our network. For instance, if the actual offered demand to the system is less than the saturation level (i.e., $\lambda < \bar{\lambda}$), then our capacity is underutilized. On the other hand, when $\lambda \geq \bar{\lambda}$, we know that we have excess offered demand. The second benefit is the fact that for $\lambda \geq \bar{\lambda}$, optimal allocations become insensitive to the offered demand λ, so that we do not need to worry about fluctuations in the input process as long as $\lambda \geq \bar{\lambda}$.

Let $\mu \leq \bar{\mu}$ be the target output. Then we determine the minimum offered demand $\lambda' \geq \mu$ required so that the target output of μ is feasible. For this, consider the following allocation LP:

$$\min\ \lambda \text{ such that} \tag{60}$$

$$\sum_{k=1}^{K} d_k p_{k,0} \geq \mu; \tag{61}$$

$$d_k \leq \sum_{j=1}^{M} \mu_{j,k}\delta_{j,k},\ k = 1,\ldots,K; \tag{62}$$

$$d_k \leq \lambda p_{0,k} + \sum_{i=1}^{K} d_i p_{i,k},\ k = 1,\ldots,K; \tag{63}$$

$$\sum_{k=1}^{K} \delta_{j,k} \leq 1,\ j = 1,\ldots,M; \tag{64}$$

$$d_k \geq 0,\ \delta_{j,k} \geq 0,\ j = 1,\ldots,M,\ k = 1,\ldots,K. \tag{65}$$

This time our objective is to allocate the servers such that the minimum offered demand is required while maintaining the desired output. Our decision variables are λ, d_k for $k = 1,\ldots,K$, and $\delta_{j,k}$ for $j = 1,\ldots,M$, $k = 1,\ldots,K$. The right-hand side of the first constraint (61) is the total amount of output required μ and the

left-hand side is the long-run departure rate from the system. So (61) simply means the throughput of the system should be at least μ. All the other constraints in this LP appear in the previous LP $(13)-(17)$ and have the same interpretations. Let the optimal solution to the above LP be given by $\lambda^*(\mu)$, d_k^* and $\{\delta_{j,k}^*\}$.

Theorem 3.4.2. *(a) A generalized round robin policy constructed as in Section 3.3.1 or Section 3.3.2, based on the offered demand $\lambda \geq \lambda^*(\mu)$ and allocations $\delta_{j,k}^*$, for all j, k, obtained from the solution of the allocation LP $(60)-(65)$ comes arbitrarily close to the target throughput μ. That is, the throughput μ^π of the generalized round robin policy π satisfies $\mu^\pi \geq \mu(1-\epsilon)$, where $0 < \epsilon < 1$.*

(b) We have $\bar{\lambda} = \lambda^*(\bar{\mu})$.

Proof. To simplify the notation, let $\tilde{\lambda} = \lambda^*(\mu)$. Let a policy π be designed as in Section 3.3.1 or Section 3.3.2 corresponding to $\tilde{\lambda}$ and ϵ. Then we have by Theorem 3.3.1 or 3.3.2 that $\mu^\pi \geq \mu^*(\tilde{\lambda})(1-\epsilon)$, where $\mu^*(\tilde{\lambda})$ is the solution to the allocation LP $(13)-(17)$. Note that d_k^* and $\{\delta_{j,k}^*\}$ from the LP $(60)-(65)$ also satisfy $(13)-(17)$. The constraint (61) implies that $\mu^*(\tilde{\lambda}) \geq \mu$, and hence that $\mu^\pi \geq \mu(1-\epsilon)$ as required. Together with Lemma 3.2.1, this proves part (a) of the theorem, and part (b) follows by the definition of $\bar{\lambda}$ and Theorem 3.4.1. □

Note that our generalized round robin policies depend on the offered demand λ. So an optimal assignment for a given λ may not be the best choice when the actual offered demand varies. In Section 3.5, we look at the sensitivity of the throughput to varying offered demand.

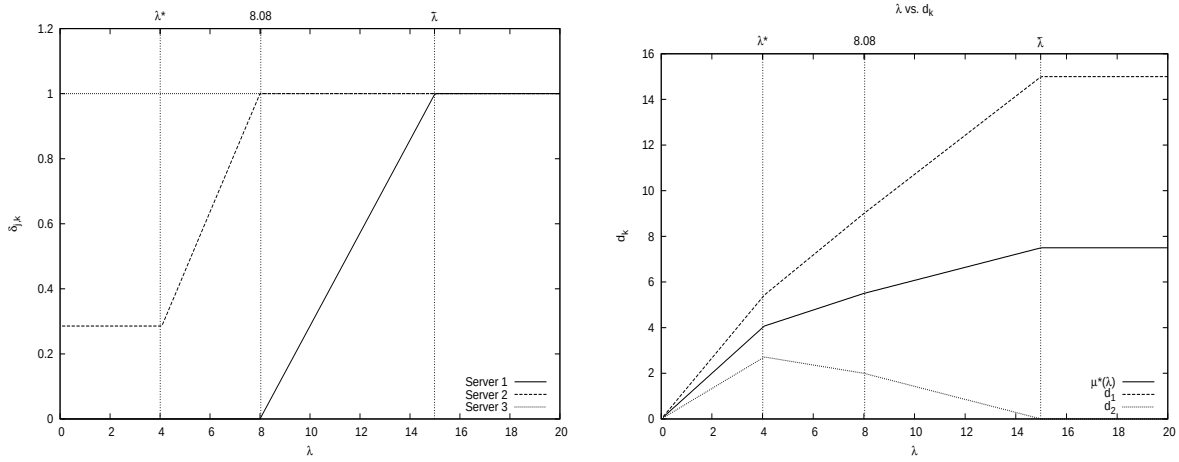
3.5 A Numerical Example

In this section, we provide in-depth analysis of an example from Section 3.2.3 (see Figure 1). Section 3.5.1 demonstrates how the optimal allocations vary as the offered demand to the system changes. Section 3.5.2 investigates the sensitivity of the optimal allocation for a given offered demand to the actual offered demand. Lastly, Section 3.5.3 simulates the same example for a given offered demand level.

3.5.1 Optimal Server Allocations Under Varying Offered Demand

In this section, we use an example to illustrate the effects on the maximum throughput of increasing the offered demand λ to the system. We will investigate the system with two classes and three servers considered earlier in Section 3.2.2 (see Figure 1

and equation (24)). Instead of looking at a single offered demand $\lambda = 6$, we consider $\lambda \in [0, 20]$ by dividing this range into 500 equal intervals and solving the allocation LP for each value of λ incrementally (i.e., $\lambda = 0.04, 0.08, \ldots, 20$). Note that for this system, we have $\lambda^* \simeq 4.0714$, $\bar{\lambda} = 15$, and $\bar{\mu} = 7.5$. Figure 2(a) gives the optimal assignments to class 1 for each server corresponding to different λ. Figure 2(b) shows d_1^*, d_2^*, and $\mu^*(\lambda)$ as a function of the offered demand λ. Note that optimal allocations for a given λ may not be unique. To avoid fluctuations in the allocations and better see the effects of instability, we consider two specific basic allocations and use them whenever they are feasible and optimal. The first specific basic allocation is obtained by solving the allocation LP given by Andradóttir, Ayhan, and Down [6]. The second specific basic solution is obtained by solving the allocation LP $(13) - (17)$ for $\lambda = \bar{\lambda}$. Then for $\lambda \leq \lambda^*$ and $\lambda \geq \bar{\lambda}$, the optimal allocations are constant and equal to the first and second specific basic solutions, respectively. When $\lambda^* < \lambda < \bar{\lambda}$, neither of the specific basic solutions is optimal, and the allocations obtained from the solution of the allocation LP $(13) - (17)$ are used.



(a) Server Assignments at Class 1 (b) λ vs. d_k

Figure 2: Optimal server assignments at class 1 and corresponding departure rates at each class as a function of λ

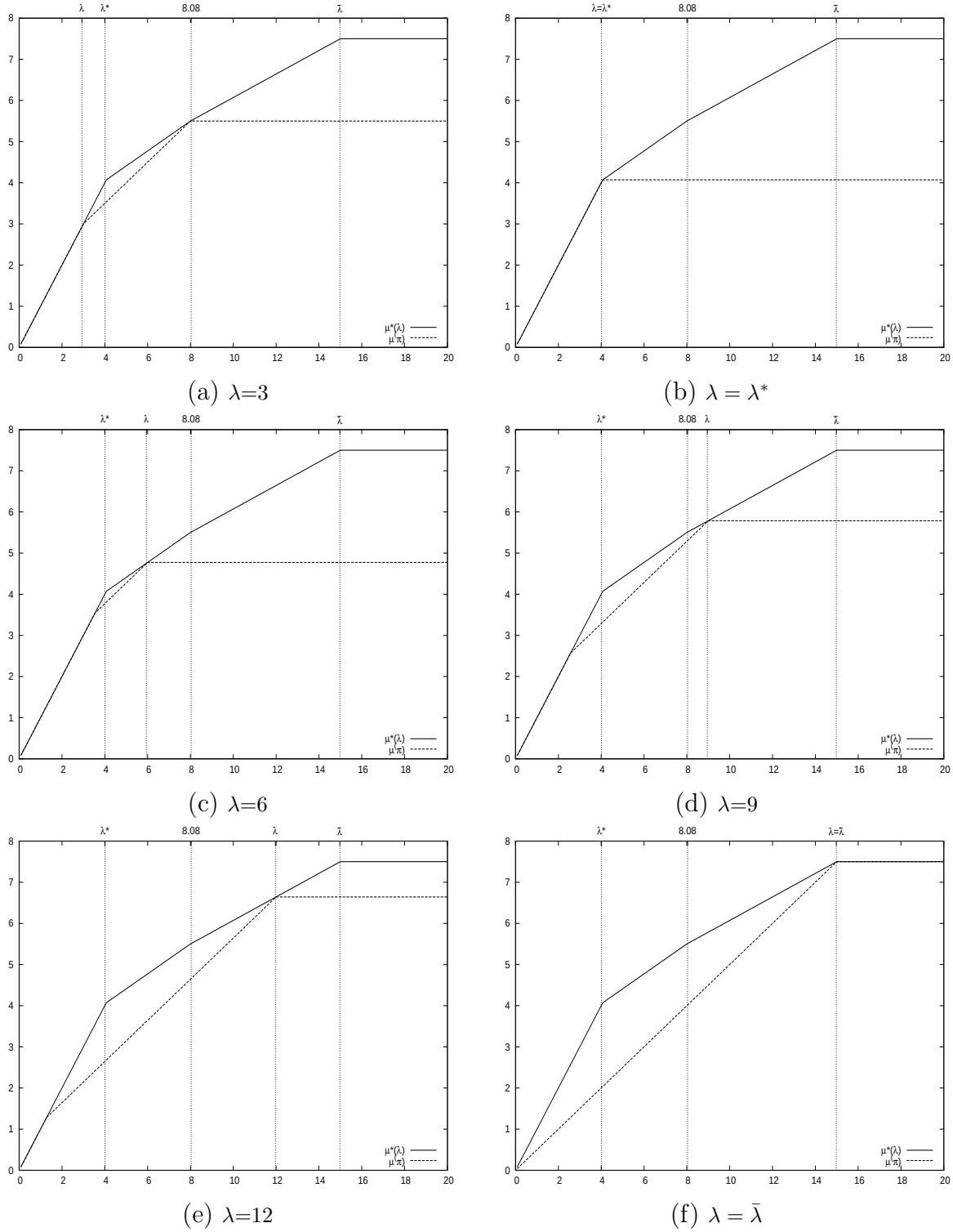
As we can see from Figure 2(a), servers 1 and 2 switch from the second class to the first class as the offered demand λ increases. Consequently, the servers prefer class 1 as long as there are customers there to process (because a customer leaving class 2 requires more service effort than one leaving class 1). But any excess capacity is devoted to class 2 since it also has an effect on the throughput. If all the servers work at class 1, then the total processing rate is 15. Hence until $\bar{\lambda} = 15$, some excess capacity is available to allocate to class 2 customers. For $\lambda \leq 8.08$, servers 2 and

3 are able to handle all of the input to class 1, with server 2 helping with class 1, increasingly with λ. After all the efforts of servers 2 and 3 are devoted to class 1 at $\lambda = 8.08$, server 1 starts to help until all of its effort is switched to class 1 as well. Figure 2(b) also shows that as expected by Lemma 3.2.1, the throughput is a piecewise-linear concave function of the offered demand level. Moreover, we observe that by allowing instability in the queueing network, it is possible for the production output to increase significantly compared to the stable throughput (in this case by a factor of almost two) given sufficient input. However, the optimal departure rates from each class d_1^* and d_2^* display different reactions to the increasing offered demand λ in parallel with optimal allocations in Figure 2(a). They both increase until server 2 starts to spend more time on the first class, so that d_2^* starts to decrease.

3.5.2 System Throughput Under Varying Offered Demand

In this section, we look at the performance of the optimal policy developed for one offered demand as a function of the actual offered demand. For this, we develop a policy based on a fixed λ, and then investigate the system performance when the actual offered demand λ' is different from λ. Figure 3 depicts the cases where the policy π is designed for $\lambda \in \{3, \lambda^*, 6, 9, 12, \bar{\lambda}\}$, respectively, and provides the optimal throughput $\mu^*(\lambda')$ and actual throughput $\mu_\lambda^\pi(\lambda')$ for different λ'. To obtain $\mu_\lambda^\pi(\lambda')$, we use the optimal fractions obtained for λ in the allocation LP (13) − (17), and solve for d_k^*, for all k, see Section 3.2.2. The actual throughput of the system differs from the optimal because the policy is designed based on the offered demand λ, and hence the assignments may no longer be optimal for another offered demand λ'. Note that in Figure 3(a), we have used the allocations obtained as a result of solving the LP (13) − (17) for $\lambda = 3$, and not the ones obtained for the point λ^*. As a result, we observe that the throughput becomes sensitive to the offered demand even for $\lambda \leq \lambda' \leq \lambda^*$. Substituting the allocations obtained at λ^* for $\lambda = 3$, Figure 3(a) would be the same as Figure 3(b).

As can be seen in Figure 3, the system performance is sensitive to the actual offered demand level. Note that λ^* is a critical point in all of the figures. Moreover, we notice a common pattern that $\mu_\lambda^\pi(\lambda')$ equals $\mu^*(\lambda')$ until some point t_1, then deviates from $\mu^*(\lambda')$, intersecting it only at a second point t_2 (if $\lambda \neq \lambda^*$), and finally becoming constant after the second intersection. For those two points t_1 and t_2, we have $0 \leq t_1 \leq \min\{\lambda, \lambda^*\}$ and $\lambda^* \leq t_2 \leq \bar{\lambda}$. Also, $\mu_\lambda^\pi(\lambda)$ is always equal to $\mu^*(\lambda)$, and in particular $t_1 = \lambda$ when $\lambda \leq \lambda^*$, and $t_2 = \min\{\lambda, \bar{\lambda}\}$ when $\lambda \geq \lambda^*$. We have two special cases, namely when $\lambda = \lambda^*$, where $t_1 = t_2 = \lambda$, and when $\lambda \geq \bar{\lambda}$, where $t_1 = 0$



(a) λ=3 (b) $\lambda = \lambda^*$

(c) λ=6 (d) λ=9

(e) λ=12 (f) $\lambda = \bar{\lambda}$

Figure 3: Sensitivity analysis when actual offered demand differs from the one designed for.

and $t_2 = \bar{\lambda}$. Also, a comparison of parts (a) and (b) of Figure 3 shows that solving the allocation LP for $\lambda = 3$, rather that $\lambda = \lambda^*$, achieves higher output for large λ'. This is because the optimal solution for $\lambda = 3$ turns out to be similar to the optimal solution for $\lambda \simeq 8$. Finally, note that the assignments are constant for $\lambda \geq \bar{\lambda}$ (see Figure 2(a)), and hence sensitivity analysis for $\lambda \geq \bar{\lambda}$ will be exactly the same as for $\lambda = \bar{\lambda}$.

If the offered demand to the system is not known beforehand, then there is no single best λ to design for, since solving for λ is not necessarily good for other λ' regardless of whether $\lambda' < \lambda$ or $\lambda' > \lambda$. However, we can still make some generalizations, since system capacity is not lost when $\lambda' < \lambda \leq \lambda^*$ and when $\lambda, \lambda' \geq \bar{\lambda}$. In particular, if the expected offered demand is less than λ^*, then it is best to design for λ^* so that no throughput is lost (see Theorem 1 in [6]). Similarly, if the expected offered demand is greater than $\bar{\lambda}$, then we design for $\bar{\lambda}$ without any loss of throughput. However, we cannot say the same when $\lambda^* < \lambda < \bar{\lambda}$. So, if the expected offered demand is between λ^* and $\bar{\lambda}$, and we design for λ, then the actual throughput cannot exceed $\mu^\pi(\lambda)$. However, we could find a value of λ that minimizes our maximum loss, which in our case corresponds to some $\lambda \in [9, 12]$, where the losses at λ^* and $\bar{\lambda}$ are equal. We could find this point using the Bisection-Extreme Point Search Algorithm (BEPSA), starting with $(\lambda^* + \bar{\lambda})/2$, then moving towards the middle point between the current solution and the extreme point (i.e., λ^* or $\bar{\lambda}$) where the difference is greater. For our case, it turns out that designing a policy for $\lambda = 11$ minimizes our loss at the extreme points.

3.5.3 Simulation Results

In this section, we give simulation results for the system analyzed in the previous subsections under an arrival stream that is a Poisson process with rate $\lambda = 6$. We assume the service requirements are exponentially distributed with mean 1 and that there are three servers whose service rates are given in the matrix H, see (24). We also assume servers switch instantaneously, so that no switching times occur. Then, from the allocation LP (13) − (17), we have $\mu^*(6) \simeq 4.7727$ and the optimum assignments are given in (25). Our aim is to observe how our allocation policy with admission and routing control (see Section 3.3.1) performs in terms of achieving the theoretical throughput value, and also to see if the sets S and U predicted by the allocation LP coincide with the ones actually observed without admission and control or controlled routing.

Next we choose ϵ =2/11 in the server assignment algorithm of Section 3.3.1, so

that $\epsilon' = 0.1$. Then server 1(3) is dedicated to class 2(1), see (25). Moreover, we have that $l_{2,1} = 35$ and $l_{2,2} = 4$, obtained from (27), satisfy step 6 of the assignment algorithm. We simulate this system for one million time units with a warm-up period of length 50,000. We divide the runtime into 40 batches for constructing a 95 percent confidence interval on the throughput of the system. We expect the throughput of the system to approach $\mu^*(6)(1-\epsilon) \simeq 3.9049$ (see Theorem 3.3.1) and all the nodes to be stable. Figure 4 shows the throughput rate $D^\pi(t)/t$ as a function of time. We observe that the throughput approaches its limiting value from above. The resulting 95 percent confidence interval for the throughput is (3.9007, 3.9101) with an average of 3.9054. Figure 5 shows the queue lengths over time at classes 1 and 2. As expected given the results of Section 3.3.1, the queue length at both classes displays stable behavior.

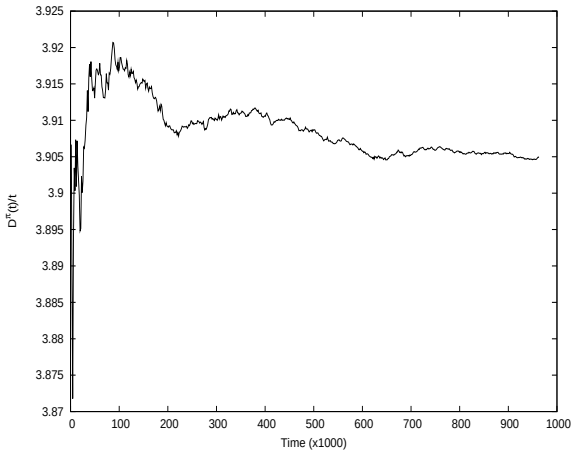
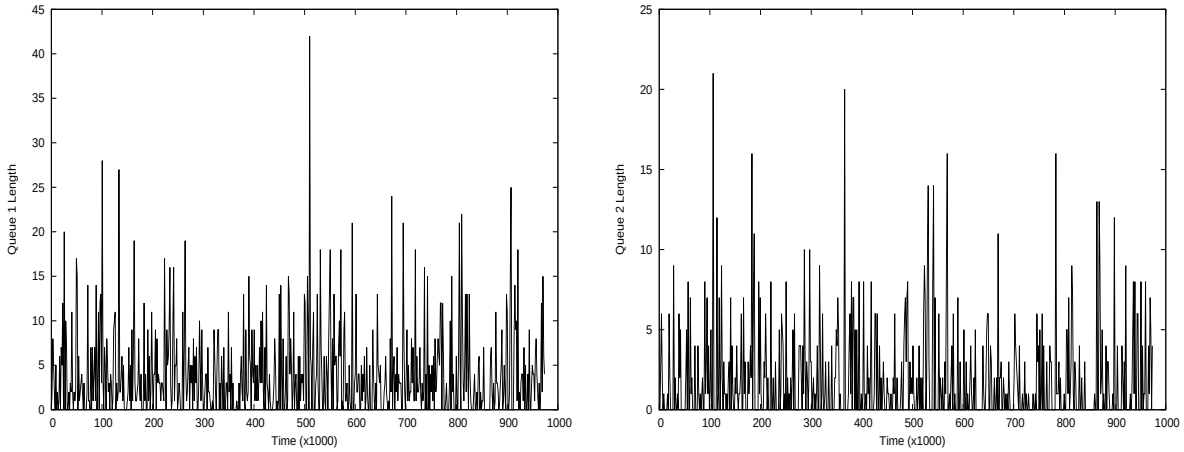


Figure 4: Average throughput with admission and routing control.

Finally, our aim is to observe the system if we the apply the policy of Section 3.3.1 without admission and routing controls. For this, we follow the same steps as in Section 3.3.1, but omit steps 2 and 3 and choose $\epsilon' = 0.1$ in step 6. Then we expect the throughput to be no smaller than $\mu^*(6)(1-\epsilon') \simeq 4.2954$. As before, we have $l_{2,1} = 35$ and $l_{2,2} = 4$, obtained from (27), satisfy step 6 of the assignment algorithm. We simulate this system for eight million time units with a warm-up period of length 300,000. A longer run length is chosen for this version of the system to observe the queue length process of class 1 (which is expected to be stable, see Section 3.2.3) for a longer period of time. We divide the run time into 40 batches for constructing a 95 percent confidence interval on the throughput of the system. Figure 6 shows the throughput rate $D^\pi(t)/t$ as a function of time. The resulting 95 percent confidence interval for the throughput is (4.7708, 4.7736) with an average of 4.7722. Figure



(a) Queue Length at Class 1 (b) Queue Length at Class 2

Figure 5: Queue lengths with admission and routing control.

7 shows the queue lengths over time at classes 1 and 2. In accordance with the results of Section 3.2.3, the queue length at class 1 displays stable behavior, whereas the queue length at class 2 increases over time. Thus the stable and unstable sets in the original queueing system operating under this policy appear to coincide with the stable and unstable sets S and U defined in (22) and (23) for the allocation LP $(13) - (17)$. As we observe, dropping steps 2 and 3 of policy of Section 3.3.1 results in significantly increased throughput at a cost of having an unstable system. This is the case because we do not reject any incoming demand (i.e., no admission control) and keep the second class busy at all times (i.e., no routing control).

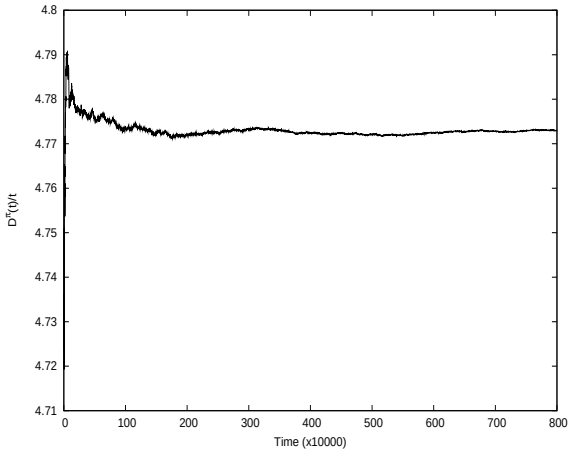
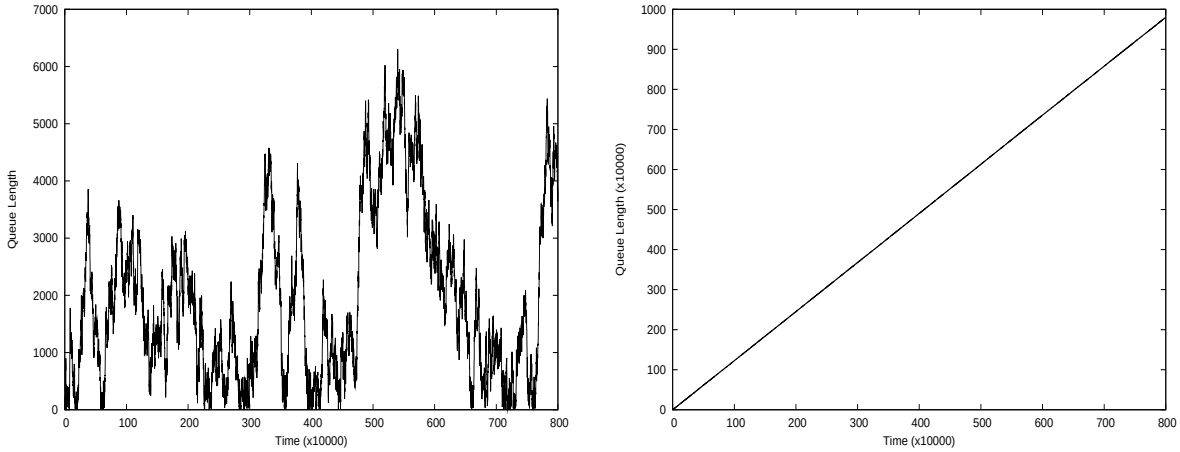


Figure 6: Average throughput without admission or routing control.



(a) Queue Length at Class 1 (b) Queue Length at Class 2

Figure 7: Queue lengths without admission or routing control.

3.6 Conclusions

We have developed generalized round robin server assignment policies for a possibly unstable queueing network with flexible servers, i.i.d. interarrival, service, and switching times, and probabilistic routing. These policies are shown to achieve any throughput less than the maximum value computed using a simple LP. In fact, allowing instability can increase the production throughput significantly given sufficient demand, resulting in higher revenues. We have also shown how to determine the saturation input and the corresponding maximum output, and provided means to check the feasibility of a desired output given the available offered demand.

One drawback for a given server assignment policy is the sensitivity of the throughput to fluctuations in the offered demand. We have shown that this sensitivity is eliminated and our policies are robust as long as the system is stable or the offered demand is above the saturation level. We have also discussed how to choose offered demand to base a policy on that minimizes the maximum loss. In actual production systems, offered demand often changes over time. In that case, we can simply modify our policies by letting the server allocations adjust with time according to the forecasted demand.

CHAPTER IV

INSPECTION LOCATION IN CAPACITY-CONSTRAINED LINES

In this chapter, we study the effects of inspection and repair stations on the production capacity and product quality in a serial line with possible inspection and repair following each operation. We consider multiple defect types and allow for possible inspection errors that are defect dependent. Unlike previous works, our analysis captures the possibility of increasing production capacity by scrapping or repairing defective items before a bottleneck operation station, and hence reducing the waste of operation capacity on defective products. Our objective is to maximize the total profit rate function that combines the effects of bottlenecks on throughput with product quality, as opposed to previous papers where the objective is either to meet minimum outgoing quality levels, or to minimize total costs, or to maximize total profit without regard to increasing the effective capacity of bottlenecks.

The organization of this chapter is as follows. In Section 4.1, our network model, assumptions, and notation are described in detail, and some limiting properties are proven. Section 4.2 introduces a probability model based on a given inspection allocation strategy, and shows how the outgoing quality level, scrap probabilities, and flow rates are calculated. Based on this probability model and flow rates, we develop the profit rate function for the system in Section 4.3, taking into account the costs incurred, as well as the revenue generated. In Section 4.4, we develop an admission control policy for a given allocation strategy that results in cost reduction, and also introduce nonlinear programs for determining the optimal inspection locations and levels when all repair stations are known to be stable. Section 4.5 provides numerical results that show how the inspection allocation decisions are determined under different parameters for a two station system. We also demonstrate that bottleneck considerations for determining the best inspection locations can lead to different inspection decisions than previous models (that do not take the capacity of the system into account). Finally, we summarize our findings in Section 4.6.

4.1 Queueing Network

In this section, we introduce the queueing network model in detail and also provide results about its asymptotic behavior. More specifically, in Section 4.1.1, the operating logic for the production network, along with the notation, are introduced. In Section 4.1.2, we show that the departure process from each server satisfies certain limiting properties.

4.1.1 Model Description

In this section, we describe the production process, assumptions, and notation in detail. Our model consists of an arbitrary number N of operation stations in tandem. After each operation station, we place an inspection station with an associated repair station. We will use the notation $O_1,\ldots,O_N$ to refer to the N consecutive operation stations, $I_1,\ldots,I_N$ to refer to the N consecutive inspection stations, and finally $R_1,\ldots,R_N$ to refer to the N consecutive repair stations. All stations have given capacities and operate under the First Come First Serve (FCFS) scheduling policy. Note that assuming that an inspection and repair station are associated with every operation station is without loss of generality because we can always remove an inspection and/or repair activity through an appropriate choice of parameter values. Finally, if the production process starts with an inspection instead of an operation, we can simply let O_1 be a dummy operation station with infinite capacity.

We have a finite set D of possible defects with $|D|$ elements. At each operation station O_i, a defect $j \in D$ could be incurred independently on different parts with some given probability $p_{i,j}$. Letting $p_{i,j} = 0$, we can turn off the possibility that defect j occurs at station O_i, so that only a subset $S_i = \{j \in D : p_{i,j} > 0\}$ of defects can occur at station O_i. Different defects $j \in S_i$ are introduced independently on the same unit. We assume that no defects are introduced at the inspection and repair stations, as well as at the dummy operation stations O_0 and O_{N+1}.

After each operation station O_i, $i = 1,\ldots,N$, units are routed to the associated inspection station I_i. Inspection station I_i might inspect only a fraction f_i of the incoming parts for some set $D_i \subseteq \bigcup_{j\le i} S_j$ of defects that are inspected for at inspection station I_i. Consequently, a complete inspection station I_i and the associated repair station R_i can be turned off by letting $f_i = 0$. Although Lindsay and Bishop [66] and Wiel and Vardeman [96] show that inspecting either all units or no units (so that $f_i \in \{0,1\}$, $\forall i$) yields minimal total goodwill and inspection cost for systems with Bernoulli product characteristics and independent defect propagation for each

item (as is the case for our model), we allow partial inspection for the following reasons. First and most importantly, we could make an inspection station faster through partial inspection when it is a bottleneck station for the system. Secondly, even when full inspection is not desirable, we can use partial inspection to stabilize a down-stream bottleneck. We do not retain information about whether a part was previously inspected for a specific defect. This is for example reasonable if inspection is not repeated for defects at later stages unless the defect can be reintroduced.

Our model allows the inspection process to be imperfect in that a product that does not have defect j when it is inspected at station I_i might be classified as having defect j with probability $\alpha_{i,j}$, which constitutes Type 1 error. Likewise, a product having defect j that is inspected for at station I_i might be classified as being nondefective with probability $\beta_{i,j}$, constituting Type 2 error. We assume that the inspection process for different defects is independent at each of the inspection stations and that the inspection process at different inspection stations is independent.

We assume that inspections are only carried out for defects that necessitate either that the part be scrapped (major defects) or repaired (minor defects). Hence $D_i = D_i^S \cup D_i^R$, where D_i^S is the set of major defects that require 'Scrapping' at inspection station I_i and D_i^R is the set of minor defects requiring 'Repair'. If a unit is classified by the inspection station to have at least one major defect, then the unit can not be repaired and is scrapped. On the other hand, if a unit is free of any major defects but has at least one minor defect, then it is routed to the associated repair station R_i. By contrast, a unit is routed to the next operation station O_{i+1} if the unit is not inspected, or if the unit passes the inspection (i.e., it is not found to have any defects). If a stage consist only of inspection without any repair, then D_i^R is empty and $D_i = D_i^S$, so that all defects are serious and defective unit are scrapped without any repair attempt. Thus, repair station R_i can be turned off by letting $D_i^R = \emptyset$. Let s_i^I represent the fraction of inspected units scrapped at I_i (i.e., the unit is classified as nonconforming in some defect $j \in D_i^S$) and r_i^I be the fraction of inspected units routed from the inspection station I_i to the repair station R_i (i.e., the part passes the inspection for all $j \in D_i^S$, but is classified as nonconforming for some defect $j \in D_i^R$). Finally, let o_i^I represent the fraction of inspected units routed from the inspection station I_i to the next operation station O_{i+1}.

We assume that for all units that are routed to repair station R_i, repair is attempted for all defects that are captured by the associated inspection station I_i. The repair process is independent for different defects and the repair probability for defect j depends on whether the unit actually has defect j or not, given by $q_{i,j}$ and

1, respectively. If the repair operation fails on any of the defects, then the part is scrapped. On the other hand, if the repair is successful on all known defects at R_i, then the part is sent to the operation station O_{i+1}. Let s_i^R be the fraction of units that are scrapped at repair station R_i (i.e., at least one of the repair operations fail) and o_i^R be the fraction of units routed to the next operation station O_{i+1}. Clearly the routing fractions satisfy

$$s_i^I + r_i^I + o_i^I = 1,$$
$$s_i^R + o_i^R = 1.$$

A graphical representation of our model is given in Figure 8 along with the rate notation used. Note that since we allow the system to be capacity constrained, the output from a given station is not necessarily equal to the input to that station. Let λ_i represent the arrival rate to operation station O_i and λ_i^O be the corresponding output rate. Similarly, let λ_i^I and λ_i^R be the output rates from the inspection and repair stations I_i and R_i, respectively. The flow rate between I_i and R_i, I_i and O_{i+1}, and R_i and O_{i+1} are denoted by λ_i^{IR}, λ_i^{IO}, and λ_i^{RO}, respectively. Also the scrap rates at the inspection and repair stations I_i and R_i are denoted by ν_i^I, and ν_i^R, respectively. Lastly, the production rates at the operation and repair stations O_i and R_i are denoted by μ_i^O and μ_i^R. However, the processing rate at inspection station I_i will be modelled by μ_i^I/f_i to emphasize the dependence on the fraction of parts inspected (so μ_i^I is the conditional service rate given the part is inspected). All processing times (operation, inspection, and repair) are assumed to be generally distributed i.i.d. sequences with finite variances. We also include dummy start and delivery nodes with station numbers O_0 and O_{N+1}, respectively, so that we can scrap before the first operation station O_1 at the rate ν_0^O as a form of admission control, and represent the system output by λ_{N+1}^O. The exogenous interarrival time sequence to the dummy operation station O_0 is assumed to be i.i.d. with general distribution and rate $\lambda = \lambda_0$. The production rates of the dummy operation stations are given by $\mu_0^O = \mu_{N+1}^O = \infty$.

In calculating the total profit rate, we need to know the flow rate into each station, which requires knowing the fraction of units with certain defect structures at various stages of the production process. Production can be either demand or capacity constrained, i.e., it could be constrained either by the arrival rate λ or by the processing rate of any of the operation, inspection, or repair stations. Although the throughput of the system can not exceed the capacity of any of the inspection or operation stations, it is not similarly restricted by the capacity of the repair stations

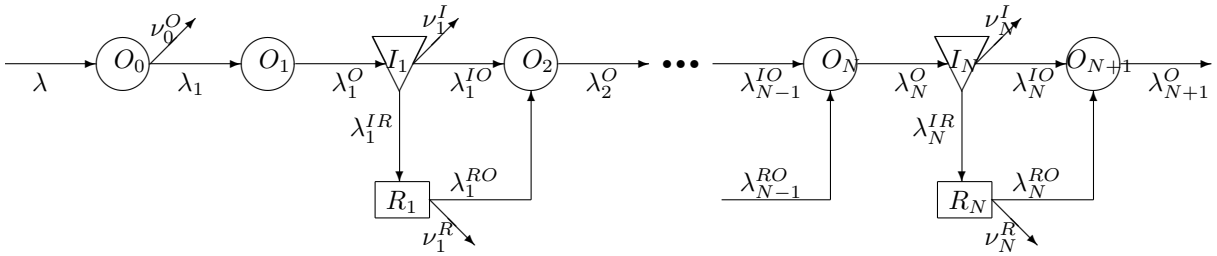


Figure 8: Model and rate notation

because of the structure of the production network. Note that when the system is capacity constrained, it may not be able to process all incoming parts.

Note that, as described above, our model with multiple defect types, error prone inspection and repair, and fractional inspection allows for more generality than the previous studies on the inspection allocation problem. Moreover, by considering throughput in the capacity constrained system and comparing profit rate functions, we account for the effects of inspection on bottleneck stations.

4.1.2 Asymptotic Properties

In this section, we define some cumulative processes for the queueing network model described in the previous section. Our aim is to show that departure processes from each server satisfy certain limiting properties. Let $A_0^O(t)$ denote the number of exogenous arrivals to the dummy operation station O_0 in $(0,t]$. For $i = 1, \ldots, N$, the processes $A_i^O(t)$, $A_i^I(t)$, and $A_i^R(t)$ are the cumulative number of jobs that arrive to operation, inspection, and repair stations O_i, I_i, and R_i, respectively, during $(0,t]$. Similarly, $N_i^O(t)$, $N_i^I(t)$, and $N_i^R(t)$ denote the cumulative number of jobs that exit operation, inspection, and repair stations O_i, I_i, and R_i, respectively, during $(0,t]$ for $i = 1, \ldots, N$.

The serial structure of the production network allows us to analyze each node sequentially. Consider some station in isolation with cumulative arrival process $A(t)$. Initially, there are $Q(0)$ jobs at the station and $Q(t)$ is the number of jobs at the station at time t. Let $B(t)$ denote the total amount of busy time for the server until time t and $S(t)$ be the potential number of service completions if the server is always busy in $(0,t]$. The actual number of departures until time t is given by $N(t)$. Then

the cumulative variables satisfy the following queueing network equations

$$\begin{aligned} Q(t) &= Q(0) + A(t) - N(t), \\ B(t) &= \int_0^t 1\{Q(s) > 0\} ds, \\ N(t) &= S(B(t)), \end{aligned} \tag{66}$$

where $1\{\cdot\}$ is the indicator function. Next we state that given that the input process satisfies certain limiting properties, so does the output process.

Proposition 4.1.1. *Consider a queue with a server that has i.i.d. processing times with rate μ, and assume that the input process $\{A(t)\}$ satisfies*

$$\lim_{t\to\infty} \frac{A(t)}{t} = \lambda \text{ almost surely (a.s.)} \tag{67}$$

Then for the output process $\{N(t)\}$, we have

$$\lim_{t\to\infty} \frac{N(t)}{t} = \min\{\lambda, \mu\} \text{ a.s.} \tag{68}$$

Proof. By the Strong Law of Large Numbers (SLLN), we have $\lim_{t\to\infty} S(t)/t = \mu$ a.s. Then, by the proof of Lemma 5.8 of Chen and Yao [25] and equation (67), we have

$$\bar{A}^n(t) = \frac{A(nt)}{n} \overset{a.s.}{\to} \lambda t \text{ as } n \to \infty, \text{ u.o.c., and } \bar{S}^n(t) = \frac{S(nt)}{n} \overset{a.s.}{\to} \mu t \text{ as } n \to \infty, \text{ u.o.c.,}$$

where u.o.c. stands for "uniformly on compact sets". Hence we can construct a fluid model for this single server queue as in Theorem 6.5 of Chen and Yao [25]. Let $\bar{Q}(t) = \lim_{n\to\infty} Q(nt)/n$ be the corresponding queue length fluid limit with initial condition $\bar{Q}(0) = 0$. When $\rho \leq 1$, $\bar{Q}(t) = 0$ u.o.c., by Chen and Yao [25], Remark 6.7, implying that $\lim_{t\to\infty} Q(t)/t = 0$ a.s. by the definition of uniform convergence on compact sets. Then it follows from (66) and (67) that $\lim_{t\to\infty} N(t)/t = \lambda$ a.s. Similarly, when $\rho > 1$, $\bar{Q}(t) = (\lambda - \mu)t$ u.o.c., implying that $\lim_{t\to\infty} Q(t)/t = \lambda - \mu$ a.s. Then (66) and (67) imply that $\lim_{t\to\infty} N(t)/t = \mu$ a.s. □

Now, for the dummy operation station, assumption (67) is satisfied because of the i.i.d. interarrival times. Then applying Theorem 4.1.1 in a recursive manner and exploiting the feedforward structure of the network shows that all departure rates exist and satisfy (68). Note that possible splits and joins at the inspection and repair stations do not complicate the analysis. For instance, at an inspection station I_i with full inspection $f_i = 1$, given that a fraction r_i^I of items are routed to repair station R_i (see Section 4.2.2), then

$$\lim_{t\to\infty} \frac{A_i^R(t)}{t} = \lim_{t\to\infty} \frac{A_i^R(t)}{N_i^I(t)} \times \frac{N_i^I(t)}{t} = r_i^I \times \min\{\lambda_i^O, \mu_i^I\} \text{ a.s.} \tag{69}$$

4.2 Defect Propagation

In this section, we derive the fraction of units with given defect structures at different stations in the network. The flow rates at various stages of the production process will then follow as described in Section 4.1.2. Our analysis will take into account that since the inspection of the units is error prone, units may be defective even after inspection or repair. We start by the analyzing the status of units leaving operation station O_i in Section 4.2.1, continue by analyzing units leaving inspection and repair stations I_i and R_i in Sections 4.2.2 and 4.2.3, respectively, and finally investigate units entering the succeeding operation station O_{i+1} in Section 4.2.4. In this way, we completely describe the i^{th} stage, and continuing N times in a similar manner, we can characterize the whole production process.

In the system analysis to follow, an inspection policy (hence f_i, $\forall i$) is assumed to be given. The analysis is started with the input rate λ to the system. Note that how much to admit to the system, hence the scrap rate ν_0^O, is a policy parameter, assumed to be known. Since $\mu_0^O = \infty$, we have $\lambda_1 = \lambda - \nu_0^O$. Later, in Section 4.4.1, we will show how to determine the best admission policy for a given inspection plan. Our analysis is initialized with the information on the incoming defect fractions $\pi_{1,j}^O$ for all j, that represent the fraction of raw materials that are already defective. Note that this is requirement is not limiting since we can simply set $\pi_{1,j}^O = 0$ for all j if the incoming defect fraction information is not available.

We start by introducing some useful notation. Define $\pi_{i,j}^O$, $\pi_{i,j}^I$, and $\pi_{i,j}^R$ as the long-run average fraction of units arriving at operation station O_i, inspection station I_i, and repair station R_i, respectively, that already have defect j. Similarly, $\pi_{i,j}^{IO}$ is the fraction of units routed from inspection station I_i to operation station O_{i+1} that has defect j. Finally, the fraction of units still having defect j after the repair operation is denoted by $\pi_{i,j}^{RO}$. Figure 9 shows the i^{th} stage and the fraction of units that have defect j at stage i after each of the steps. We discuss how $\pi_{i,j}^I$, $\pi_{i,j}^{IO}$, $\pi_{i,j}^R$, $\pi_{i,j}^{RO}$, and $\pi_{i,j}^O$ are computed next in Sections 4.2.1 through 4.2.4 below. Note that when division by zero occurs, it is easy to see that the corresponding numerator is also zero. To avoid these trivial cases, we adopt the convention $0/0 = 0$.

4.2.1 Departures from Operation Stations

At this point in the calculations, we already know all the information for the stations in stages $1, \ldots, i-1$, as well as the arrival rate λ_i, and the fraction of parts that have defect $j \in D$, before operation station O_i, given by $\pi_{i,j}^O$. At operation station O_i, a

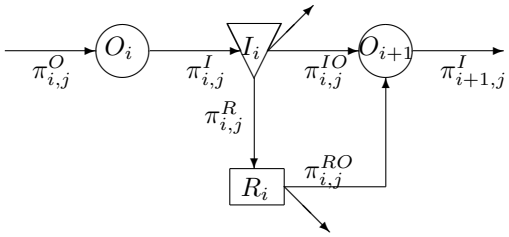


Figure 9: Notation for the fraction of units with defect j at the i^{th} stage

unit could acquire any defect $j \in S_i$ with probability $p_{i,j}$; the unit is unaltered in defect probability distributions for all defects $j \notin S_i$. Consequently, we have

$$\pi^I_{i,j} = \pi^O_{i,j} + p_{i,j}(1 - \pi^O_{i,j}), \text{ for } i = 1, \ldots, N, j \in D, \tag{70}$$

$$\lambda^O_i = \min(\lambda_i, \mu^O_i). \tag{71}$$

The second equality follows from Proposition 4.1.1. To derive the first equality, let $A^O_{i,j}(t)$, $A^I_{i,j}(t)$, and $A^R_{i,j}(t)$ be the total number of units arriving at stations O_i, I_i, and R_i, respectively, in $(0, t]$ that have defect $j \in D$. Similarly, $N^O_{i,j}(t)$, $N^I_{i,j}(t)$, and $N^R_{i,j}(t)$ are the total number of units with defect j that depart stations O_i, I_i, and R_i, respectively, until time t. Then by definition we have $\pi^O_{i,j} = \lim_{t\to\infty} A^O_{i,j}(t)/A^O_i(t)$. We have

$$\begin{aligned}\pi^I_{i,j} &= \lim_{t\to\infty} \frac{A^I_{i,j}(t)}{A^I_i(t)} = \lim_{t\to\infty} \frac{N^O_{i,j}(t)}{N^O_i(t)} = \lim_{t\to\infty} \frac{A^O_{i,j}(t)}{A^O_i(t)} + \lim_{t\to\infty} \frac{A^O_i(t) - A^O_{i,j}(t)}{A^O_i(t)} p_{i,j} \\ &= \pi^O_{i,j} + p_{i,j}(1 - \pi^O_{i,j}),\end{aligned}$$

where the third equality takes both units that already have defect j and units that acquire defect j at operation station O_i into account, as well as the FCFS service discipline and the i.i.d. assumption on the introduction of failures. Similar arguments will be used below (see, e.g., equation (73)), without detailed explanation.

4.2.2 Departures from Inspection Stations

Next we analyze the status of units departing from inspection station I_i. At this point, we already know the fractions $\pi^I_{i,j}$, $j \in D$, and the input rate λ^O_i to I_i. Let $\hat{N}^I_i(t)$ denote the total number of departing units inspected at I_i, so that $f_i = \lim_{t\to\infty} \hat{N}^I_i(t)/N^I_i(t)$. To make the analysis easier, we define the vector $W_i = (W_{i,1}, W_{i,2}, \ldots, W_{i,|D|})$ that holds information on the current defect state for a unit arriving

at inspection station I_i, so that

$$W_{i,j} = \begin{cases} 1 & \text{if the unit arriving at } I_i \text{ has defect } j, \\ 0 & \text{otherwise.} \end{cases}$$

It is possible for the inspection station I_i to make classification errors on each defect $j \in D_i$. Consequently, we let $E_i = (E_{i,1}, E_{i,2}, \ldots, E_{i,|D|})$ be a vector that holds information on whether the inspection station I_i made an inspection error for the defects $j \in D$, so that

$$E_{i,j} = \begin{cases} 1 & \text{if there is an inspection error on defect } j \in D \text{ at station } I_i, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\tilde{P}(W_i = w)$ represent the long-run average fraction of parts having the particular defect structure w. We will use the tilde notation whenever we refer to such fractions. We have assumed that $(W_{i,j}, E_{i,j})$ are independent for different $j \in D$ (see Section 4.1.1). We have

$$\tilde{P}(W_i = w) = \prod_{j \in D} \tilde{P}(W_{i,j} = w_j), \tag{72}$$

where $\tilde{P}(W_{i,j} = 1) = \pi^I_{i,j}$ and $\tilde{P}(W_{i,j} = 0) = 1 - \pi^I_{i,j}$ a.s. To see this, consider a case with $|D| = 2$, and let us obtain the fraction of items with a particular defect structure $W_{i,1}, W_{i,2}$. Now, in the whole population, a fraction $\pi^I_{i,1}$ will have $W_{i,1} = 1$. By independence, among those items, a fraction $\pi^I_{i,2}$ will satisfy $W_{i,2} = 1$. Hence the fraction of items with both defects is given by the product $\pi^I_{i,1} \times \pi^I_{i,2}$. Similar arguments are used to derive other needed quantities (see, e.g., equation (74)) without detailed explanation.

The probability of Type 1 and Type 2 errors as well as the fraction of units having the particular inspection event is given by

$$\begin{aligned} \alpha_{i,j} &= P(E_{i,j} = 1 | W_{i,j} = 0) = \tilde{P}(E_{i,j} = 1 | W_{i,j} = 0),\ j \in D_i; \\ \beta_{i,j} &= P(E_{i,j} = 1 | W_{i,j} = 1) = \tilde{P}(E_{i,j} = 1 | W_{i,j} = 1),\ j \in D_i; \\ 1 &= P(E_{i,j} = 0 | W_{i,j}) = \tilde{P}(E_{i,j} = 0 | W_{i,j}),\ j \in D \setminus D_i. \end{aligned}$$

Also, by the independence properties of the inspection process, the fraction of units having particular defect structures and inspection events is given by

$$\tilde{P}(W_i = w, E_i = e) = \tilde{P}\{W_i = w\} \prod_{j \in D} \tilde{P}(E_{i,j} = e_j | W_{i,j} = w_j),\ \forall w, e \in \{0,1\}^{|D|}.$$

Next we find the routing rates from inspection station I_i. For this, let $d_{i,j}$ be the fraction of units that inspection station I_i classifies as nonconforming in defect j given that it is inspected. Then we have

$$d_{i,j} = \begin{cases} \pi^I_{i,j}(1-\beta_{i,j}) + (1-\pi^I_{i,j})\alpha_{i,j} & j \in D_i, \\ 0 & \text{otherwise.} \end{cases} \tag{73}$$

An inspected unit is scrapped when it is classified as having at least one of the defects $j \in D^S_i$. Therefore, the fraction of inspected units that are scrapped at inspection station I_i is given by subtracting those classified as defect free from the whole population, i.e.,

$$s^I_i = 1 - \prod_{j \in D^S_i} (1 - d_{i,j}). \tag{74}$$

Among those units that pass all defects $j \in D^S_i$, the ones classified as having at least one defect $j \in D^R_i$ would be routed to the repair station R_i, so that the total fraction routed to R_i is given by

$$r^I_i = [1 - s^I_i][1 - \prod_{j \in D^R_i} (1 - d_{i,j})]. \tag{75}$$

Finally, the following fraction of inspected parts is routed to the next operation station O_{i+1} from I_i

$$o^I_i = 1 - r^I_i - s^I_i. \tag{76}$$

Since not all units are inspected, general routing fractions out of I_i can be calculated as follows

$$\begin{aligned} \tilde{P}(\text{unit is routed from } I_i \text{ to } O_{i+1}) &= 1 - f_i(1 - o^I_i) \text{ a.s.}, \\ \tilde{P}(\text{unit is routed from } I_i \text{ to } R_i) &= f_i r^I_i \text{ a.s.}, \\ \tilde{P}(\text{unit is scrapped at } I_i) &= f_i s^I_i \text{ a.s.}, \end{aligned}$$

resulting in the following flow rates out of inspection station I_i

$$\lambda^I_i = \min(\lambda^O_i, \mu^I_i / f_i), \tag{77}$$

$$\lambda^{IO}_i = \lambda^I_i [1 - f_i(1 - o^I_i)], \tag{78}$$

$$\lambda^{IR}_i = \lambda^I_i f_i r^I_i, \tag{79}$$

$$\nu^I_i = \lambda^I_i f_i s^I_i. \tag{80}$$

After inspection station I_i, those units that are routed to repair station R_i, a fraction $\pi^R_{i,j}$ of them will have defect $j \in D$. For $j \in D^S_i$, this happens when the unit is selected for inspection, has an undetected defect $j \in D^S_i$, but is sent to repair station R_i for some defect $j \in D^R_i$. For $j \in D^R_i$, this happens when a unit which actually has defect $j \in D^R_i$ is selected for inspection, passes the inspection for all $j \in D^S_i$, but either fails for defect $j \in D^R_i$ (i.e., no inspection error), or passes for the defect $j \in D^R_i$ (i.e., inspection error) but is sent to the repair station for some other defect $k \in D^R_i \setminus \{j\}$. Finally, a unit might also have defect $j \in D \setminus D_i$, when such a unit is selected for inspection, passes for all $j \in D^S_i$, but fails for some $k \in D^R_i$. Then, to calculate $\pi^R_{i,j}$, we need to know the fraction of units that are sent to repair station R_i for some defect other than $j \in D^R_i$, which will be represented by $r^I_{i,j}$, and is given by

$$r^I_{i,j} = \begin{cases} \prod_{k \in D^S_i}(1-d_{i,k})[1-\prod_{k \in D^R_i \setminus \{j\}}(1-d_{i,k})] & j \in D^R_i, \\ \prod_{k \in D^S_i \setminus \{j\}}(1-d_{i,k})[1-\prod_{k \in D^R_i}(1-d_{i,k})] & j \in D^S_i, \\ r^I_i & j \in D \setminus D_i. \end{cases} \tag{81}$$

Then the fraction of units that have defect $j \in D$, $\pi^R_{i,j}$, out of all the units that are routed to repair station R_i, is given by

$$\pi^R_{i,j} = \begin{cases} \dfrac{\pi^I_{i,j}(1-\beta_{i,j})(1-s^I_i) + \pi^I_{i,j}\beta_{i,j}r^I_{i,j}}{r^I_i} & j \in D^R_i, \\ \dfrac{\pi^I_{i,j}\beta_{i,j}r^I_{i,j}}{r^I_i} & j \in D^S_i, \\ \pi^I_{i,j} & \text{otherwise.} \end{cases} \tag{82}$$

The ratios in the above equalities follow because when taking the limits, we divide the total number of units with the required characteristics by the total number of units routed to repair station R_i (similar arguments are used elsewhere, e.g., (83)).

A fraction $\pi^{IO}_{i,j}$ of units that are sent to the following operation station O_{i+1} directly from inspection station I_i might also have defect $j \in D$. This happens for $j \in D_i$ when a unit that actually has defect $j \in D_i$ is not selected for inspection, or when it is selected but passes the inspection for all $k \in D_i$ (i.e., there is an inspection error). For defects $j \in D \setminus D_i$, this fraction does not depend on whether the unit was selected for inspection. Therefore, among all units that are routed to operation station O_{i+1}, the fraction that will have defect $j \in D$ is given by

$$\pi^{IO}_{i,j} = \begin{cases} \dfrac{\pi^I_{i,j}(1-f_i) + \pi^I_{i,j}f_i\beta_{i,j}\prod_{k \in D_i \setminus \{j\}}(1-d_{i,k})}{o^I_i} & j \in D_i, \\ \pi^I_{i,j} & \text{otherwise.} \end{cases} \tag{83}$$

4.2.3 Departures from Repair Stations

Our next analysis involves units that are routed to repair station R_i from inspection station I_i. At this point, we already have information on the input rate λ_i^{IR} to repair station R_i and the fraction of units with defect j, $\pi_{i,j}^R$, $\forall j$. Since the repair probability depends on whether the unit is actually defective, we need to retain information on whether there has been a classification error for any defect that fails inspection. However, since repair is not attempted for defects passing inspection, we do not need to know whether or not the unit actually has those defects. Let F_0 and F_1 denote the outcome that a unit fails inspection for a particular defect but is actually nondefective or defective, respectively, and let P be the outcome that the unit passes the inspection for the defect. Then we have $3^{|D_i^R|} - 1$ possible repair configurations for a unit at a repair station. Let the set of these configurations at repair station R_i be denoted by Z_i. For instance, if four defects are inspected for with two of them considered serious and the other two minor, then the set Z_i has the eight elements shown in Table 1.

Table 1: Defect classifications for a unit arriving at R_i with $|D_i| = 4$ and $|D_i^S| = |D_i^R| = 2$.

$j \in D_i^S$	$j \in D_i^R$
PP	PF_0
PP	PF_1
PP	F_0P
PP	F_1P
PP	F_0F_0
PP	F_0F_1
PP	F_1F_0
PP	F_1F_1

Let $z(j)$ denote the event status for defect j in the element $z \in Z_i$. Note that

$$\begin{aligned} \tilde{P}(z(j) = P) &= \tilde{P}(W_{i,j} = 0, E_{i,j} = 0) + \tilde{P}(W_{i,j} = 1, E_{i,j} = 1) \\ &= (1 - \pi_{i,j}^I)(1 - \alpha_{i,j}) + \pi_{i,j}^I \beta_{i,j} \text{ a.s.} \end{aligned} \tag{84}$$

Similarly, for $j \in D_i^R$, we have

$$\tilde{P}(z(j) = F_0) = \tilde{P}(W_{i,j} = 0, E_{i,j} = 1) = (1 - \pi_{i,j}^I)\alpha_{i,j} \text{ a.s.}, \tag{85}$$

$$\tilde{P}(z(j) = F_1) = \tilde{P}(W_{i,j} = 1, E_{i,j} = 0) = \pi_{i,j}^I(1 - \beta_{i,j}) \text{ a.s.} \tag{86}$$

Since the inspection process is independent for different defects, we can calculate the occurrence frequency of any element z of the set Z_i by multiplying the appropriate

fractions for the defects $j \in D_i$. For instance, for the units inspected at I_i, the fraction for the first element in Table 1, $z = PPPF_0$, can be calculated as

$$\begin{aligned}\tilde{P}(z) = \tilde{P}(z = PPPF_0) &= [(1-\pi^I_{i,1})(1-\alpha_{i,1}) + \pi^I_{i,1}\beta_{i,1}][(1-\pi^I_{i,2})(1-\alpha_{i,2}) + \pi^I_{i,2}\beta_{i,2}] \\ &\quad \times[(1-\pi^I_{i,3})(1-\alpha_{i,3}) + \pi^I_{i,3}\beta_{i,3}][(1-\pi^I_{i,4})\alpha_{i,4}] \text{ a.s.}\end{aligned}$$

Let $1_{\{.\}}$ denote the indicator function. Since the repair process on all the defects are independent, we can calculate the fraction of units s^R_i that are scrapped at the repair station as follows

$$s^R_i = \frac{\sum_{z \in Z_i} \tilde{P}(z)[1 - \prod_{j \in D^R_i} q_{i,j}(z)]}{r^I_i}, \tag{87}$$

where $q_{i,j}(z) = (q_{i,j}1_{\{z(j)=F_1\}} + 1_{\{z(j)=F_0\}} + 1_{\{z(j)=P\}})$ is the repair probability for defect $j \in D^R_i$ corresponding to element $z \in Z_i$. The fraction of units routed from repair station R_i to the following operation O_{i+1} is simply given by $o^R_i = 1 - s^R_i$. Then, the rates out of repair station R_i are given by

$$\begin{aligned}\lambda^R_i &= \min(\lambda^{IR}_i, \mu^R_i), &(88)\\ \lambda^{RO}_i &= \lambda^R_i o^R_i, &(89)\\ \nu^R_i &= \lambda^R_i s^R_i.\end{aligned}$$

Next we calculate the defect fractions $\pi^{RO}_{i,j}$ of units that are routed from repair station R_i to the following operation station O_{i+1}. Even if a unit is successfully repaired, they might still be defective due to undetected defects. More specifically, for $j \in D_i$, this happens when a unit has defect j, passes the inspection for all defects in D^S_i and for defect j (i.e., inspection error), but fails for some defect $k \in D^R_i$ (and is hence sent to the repair station R_i for defect k), and the repair activity on all such defects $k \in D^R_i$ is successful (so that the unit is sent to the next operation station O_{i+1}). Let the set $Z_{i,j} \subseteq Z_i$ denote the instances in Z_i such that the unit has defect j but passes the inspection for defect j, and P_1 denote the corresponding outcome. For instance for the example of Table 1, $z_{i,3} = \{PPP_1F_0, PPP_1F_1\}$. Moreover, $\tilde{P}(z(j) = P_1) = \pi^I_{i,j}\beta_{i,j}$ and $\tilde{P}(z(j))$ for $z(j) = P$, F_0, and F_1 are calculated as before. We can calculate the occurrence frequency of any element of the set $Z_{i,j}$ by multiplying the appropriate fractions for each $j \in D_i$. For instance,

$$\begin{aligned}&\tilde{P}(z = PPP_1F_0) = \\ &\quad [(1-\pi^I_{i,1})(1-\alpha_{i,1}) + \pi^I_{i,1}\beta_{i,1}][(1-\pi^I_{i,2})(1-\alpha_{i,2}) + \pi^I_{i,2}\beta_{i,2}][\pi^I_{i,3}\beta_{i,3}][(1-\pi^I_{i,4})\alpha_{i,4}].\end{aligned}$$

The fraction of units that are routed from repair station R_i to the operation station O_{i+1} and have defect j is given by

$$\pi_{i,j}^{RO} = \begin{cases} \dfrac{\sum_{z \in Z_{i,j}} \tilde{P}(z) \prod_{k \in D_i^R} (q_{i,k} 1_{\{z(k)=F_1\}} + 1_{\{z(k)=F_0\}} + 1_{\{z(k)=P_1\}})}{r_i^I o_i^R} & j \in D_i, \\ \pi_{i,j}^I & j \in D \setminus D_i. \end{cases} \tag{90}$$

4.2.4 Arrivals to Operation Stations

In this section, we characterize the units arriving at operation station O_{i+1}. At this point, we already know the fraction of units $\pi_{i,j}^{IO}$ and $\pi_{i,j}^{RO}$ that are incoming from inspection station I_i or repair station R_i and have defect j in Sections 4.2.2 and 4.2.3, respectively. We also know the flow rates λ_i^{IO} and λ_i^{RO} into operation station O_{i+1} from I_i and R_i, respectively. Then $\pi_{i+1,j}^O$ is simply a weighted average given by

$$\pi_{i+1,j}^O = \begin{cases} \dfrac{\lambda_i^{IO} \pi_{i,j}^{IO} + \lambda_i^{RO} \pi_{i,j}^{RO}}{\lambda_i^{IO} + \lambda_i^{RO}} & j \in D_i, \\ \pi_{i,j}^I & j \in D \setminus D_i, \end{cases} \tag{91}$$

This follows because when taking the limits, we divide the total number of defective units from both I_i and R_i by the total number of units arriving at O_{i+1}. Also, the total flow rate into O_{i+1}, needed to start the $(i+1)^{th}$ stage analysis, is given by

$$\lambda_{i+1} = \lambda_{i,j}^{IO} + \lambda_{i,j}^{RO}.$$

4.3 Throughput Analysis and Cost Figures

In this section, we analyze the cost structure of the inspection model developed in Section 4.2, with the objective of comparing different scenarios on the basis of profit per unit time. Inspection and repair activities incur costs for the production system in addition to the production and goodwill costs. Let $E_i^O(t)$, $E_i^I(t)$ and $E_i^R(t)$ denote the cumulative production, inspection, and repair costs for stations O_i, I_i, and R_i during $(0, t]$, respectively. Scrapping units at station I_i and R_i results in cumulative costs $S_i^I(t)$ and $S_i^R(t)$ until time t. At the end of the production process, let $T_R(t)$ and $E_G(t)$ be the cumulative revenue and goodwill cost generated until time t. Then the total profit $T_P(t)$ is the remaining revenue after accounting for all costs involved, given by

$$T_P(t) = T_R(t) - E_G(t) - \sum_{i=1}^{N} \Big[E_i^O(t) + E_i^I(t) + S_i^I(t) + E_i^R(t) + S_i^R(t) \Big]. \tag{92}$$

We are interested in the long-run average profit per unit time for the system, $T_P = \lim_{t\to\infty} T_P(t)/t$. Next we look at the different costs in detail. We start with inspection station I_i in Section 4.3.1, and continue with repair station R_i in Section 4.3.2. Finally, we obtain the total profit for the system in Section 4.3.3.

4.3.1 Inspection Cost Computation

In this section, we derive the inspection and scrap costs incurred at inspection station I_i. Note that inspection cost for each item might be item-dependent, even though all items are inspected for the same set of defects at I_i, and it depends on the inspection policy. One good policy is to first inspect for defects in D_i^S that require scrapping, and scrap the item as soon as one such defect is found. The policy might order the defects in D_i^S based on inspection cost and occurrence frequency. For a given policy, inspecting the n^{th} unit at inspection station I_i incurs random inspection cost $X_i^I(n) \geq 0$. We assume that the sequence $\{X_i^I(n)\}$ is i.i.d. with finite mean $E[X_i^I(n)] = E_i^I$. This means that $E_i^I(t) = \sum_{n=1}^{\hat{N}_i^I(t)} X_i^I(n)$. Hence the unit time inspection cost is given by

$$\lim_{t\to\infty} \frac{E_i^I(t)}{t} = \lim_{t\to\infty} \frac{E_i^I(t)}{\hat{N}_i^I(t)} \times \frac{\hat{N}_i^I(t)}{N_i^I(t)} \times \frac{N_i^I(t)}{t} = E_i^I f_i \lambda_i^I \text{ a.s.}$$

For a given policy, disposing of the n^{th} unit at inspection station I_i incurs a possibly random scrap cost $U_i^I(n) \geq 0$. We assume that the sequence $\{U_i^I(n)\}$ is i.i.d. with finite mean $E[U_i^I(n)] = U_i^I$. The scrap costs $U_1^I \ldots U_N^I$ include any auxiliary costs related with disposing the unit. Hence one would generally expect that $U_1^I \leq \ldots \leq U_N^I$. Let $\hat{N}_i^{IS}(t)$ be the total number of units scrapped at I_i until time t. Then, in terms of the cumulative processes, this means

$$\lim_{t\to\infty} \frac{S_i^I(t)}{t} = U_i^I \lim_{t\to\infty} \frac{\hat{N}_i^{IS}(t)}{\hat{N}_i^I(t)} \times \frac{\hat{N}_i^I(t)}{N_i^I(t)} \times \frac{N_i^I(t)}{t} = U_i^I s_i^I f_i \lambda_i^I = U_i^I \nu_i^I \text{ a.s.}$$

Then the total limiting inspection and scrap costs per unit time, E_I and S_I, for the whole system are

$$E_I = \sum_{i=1}^{N} E_i^I f_i \lambda_i^I \text{ and } S_I = \sum_{i=1}^{N} U_i^I \nu_i^I. \tag{93}$$

For example, a simple policy could be to inspect for all defects in D_i^S until a detect is defected, and then if the unit passes we inspect for all the defects in D_i^R. Let the constant $H_{i,j}$ be the inspection cost per unit for inspecting a unit for defect j

at inspection station I_i. $H_{i,j}$ might inversely depend on the inspection errors $\alpha_{i,j}$ and $\beta_{i,j}$ to reflect the cost of inspection quality (the lower the inspection error values, the higher the inspection cost). Then the limiting inspection cost at I_i is given by

$$E_i^I = \sum_{j \in D_i^S} \prod_{\substack{k \in D_i^S \\ k<j}} (1 - d_{i,k}) H_{i,j} + (1 - s_i^I) \sum_{j \in D_i^R} H_{i,j}.$$

4.3.2 Repair Cost Computation

In this section, we derive repair and scrap costs incurred at repair station R_i. All units arriving at the repair station incur the repair cost since repair is attempted for all of them. Note that the cost of repairing a unit depends on defect classification information, and may also depend on the particular repair policy selected. One good policy is to order the defects in D_i^R based on repair cost and repair success probability, and scrap units as soon as repair for a particular defect is unsuccessful. For a given policy, repairing the n^{th} unit with defect information $z \in Z_i$ at repair station R_i incurs random repair cost $X_i^R(n|z) \geq 0$. We assume that the sequence $\{X_i^R(n|z)\}$ is i.i.d. with finite mean $E[X_i^R(n|z)] = E_i^R(z)$ for all $z \in Z_i$. From Section 4.2.3, we know the fraction of items $\tilde{P}(z)$ with particular defect information $z \in Z_i$. Then we obtain

$$\lim_{t \to \infty} \frac{E_i^R(t)}{t} = \lambda_i^R \sum_{z \in Z_i} \tilde{P}(z) E_i^R(z) \text{ a.s.}$$

For a given policy, disposing of the n^{th} unit at repair station R_i incurs random scrap cost $U_i^R(n) \geq 0$. We assume that the sequence $\{U_i^R(n)\}$ is i.i.d. with finite mean $E[U_i^R(n)] = U_i^R$. Then, we get

$$\lim_{t \to \infty} \frac{S_i^R(t)}{t} = U_i^R s_i^R \lambda_i^R = U_i^R \nu_i^R \text{ a.s.}$$

Then the total limiting repair and scrap costs E_R and S_R for the whole system are given by

$$E_R = \sum_{i=1}^{N} \lambda_i^R \sum_{z \in Z_i} \tilde{P}(z) E_i^R(z) \text{ and } S_R = \sum_{i=1}^{N} U_i^R \nu_i^R. \tag{94}$$

For example, a simple policy could be to attempt repair for all defects in D_i^R until the first failed repair. Repair cost might depend on both defect types and the repair station. Let $C_{i,j}$ be the unit repair cost at repair station R_i for defect j whenever the unit is classified to be nonconforming in defect j by the i^{th} inspection station and the

unit is actually defective. Similarly, $C'_{i,j}$ is the unit repair cost at repair station R_i for defect j whenever the unit is classified to have defect j but the unit actually is not defective. Then we have

$$\{\text{Repair Cost for defect } j \text{ at } R_i \text{ per unit} | W_{i,j} = w_j, E_{i,j} = e_j\} = \begin{cases} 0 & \text{if } (W_{i,j}, E_{i,j}) = (0,0), (1,1), \\ C_{i,j} & \text{if } (W_{i,j}, E_{i,j}) = (1,0), \\ C'_{i,j} & \text{if } (W_{i,j}, E_{i,j}) = (0,1), \end{cases}$$

so that limiting repair cost at R_i is given by

$$E_i^R(z) = \sum_{j \in D_i^R} \prod_{\substack{k \in D_i^R \\ k<j}} q_{i,k}(z)(1 - q_{i,j}(z)) \sum_{\substack{l \in D_i^R \\ l \leq j}} (C_{i,l} 1_{\{z(l)=F_1\}} + C'_{i,l} 1_{\{z(l)=F_0\}}).$$

4.3.3 Total Profit Computation

In this section, we derive total revenue, production cost, and goodwill cost per unit time, and also compute the total profit rate. Note that the fraction of departing units that will have defect j at station O_{N+1}, $\pi^O_{N+1,j}$, as well as the throughput λ_{N+1} of the system, are determined in Section 4.2. For the revenue calculations, the n^{th} departing unit from station O_{N+1} incurs random revenue $X^O_{N+1}(n)$. We assume that the sequence $\{X^O_{N+1}(n)\}$ is i.i.d. with finite mean $E[X^O_{N+1}(n)] = R$. Then the revenue per unit time is given by

$$\lim_{t \to \infty} \frac{T_R(t)}{t} = \lambda_{N+1} R \text{ a.s.} \tag{95}$$

Processing the n^{th} unit at operation station O_i incurs random operation cost $X^O_i(n) \geq 0$, including raw material cost. We assume that the sequence $\{X^O_i(n)\}$ is i.i.d. with finite mean $E[X^O_i(n)] = E^O_i$. This means that $E^O_i(t) = \sum_{n=1}^{N^O_i(t)} X^O_i(n)$. Hence the unit time operation cost is given by

$$\lim_{t \to \infty} \frac{E^O_i(t)}{t} = \lim_{t \to \infty} \frac{E^O_i(t)}{N^O_i(t)} \times \frac{N^O_i(t)}{t} = E^O_i \lambda^O_i \text{ a.s.} \tag{96}$$

Also, let E_O represent the total limiting operation cost throughout the production line so that

$$E_O = \sum_{i=1}^{N} E^O_i \lambda^O_i. \tag{97}$$

Goodwill cost depends on the combination of defects and final recipient's quality perception. For instance, a minor defect along with a major defect might add no or little cost to the total goodwill cost. At the same time, a defect perceived to be minor for one customer might be major for another. Let $X^G(n|w) \geq 0$ represent the random goodwill cost associated with the n^{th} completed unit having defect structure $W_{N+1} = w$ for all $w \in \{0,1\}^{|D|}$. We assume that the sequence $\{X^G(n|w)\}$ is i.i.d. with mean $E[X^G(n|w)] = E_G(w)$ for all $w \in \{0,1\}^{|D|}$. As in Section 4.2.2, we can obtain the fraction of items $\tilde{P}(W_{N+1} = w)$ at station O_{N+1} with a particular defect structure w. Finally, by a similar analysis as in the previous section, we obtain the limiting goodwill cost

$$E_G = \lim_{t\to\infty} \frac{E_G(t)}{t} = \lambda_{N+1} \sum_{w\in\{0,1\}^{|D|}} \tilde{P}(W_{N+1} = w) E_G(w). \tag{98}$$

For example, a simple model for calculating the goodwill cost is the additive model, where each defect $j \in D$ is associated with an expected goodwill cost G_j, so that

$$E_G(w) = \sum_{j\in D} \pi^O_{N+1,j} G_j.$$

For the general model, combining the results from equations (92) – (94) and (95) – (98), we obtain the limiting total profit per unit time

$$T_P \;=\; \lambda_{N+1}R - E_O - E_I - S_I - E_R - S_R - E_G \text{ a.s.} \tag{99}$$

4.4 Inspection Location and Admission Control

In real-life production lines, manufacturing and inspection operations can require considerably different amounts of time. For instance, in traditional manufacturing processes such as automobile assembly lines, manufacturing operations take much longer than inspections, because the inspection activity is simpler and consists of activities like visual inspection or simple functionality check. For example, laser and vision-based inspection systems can supply huge amounts of data about form, fit, and contour in only a few minutes (see Tolinski [89]). In such cases, the cycle time of products would be determined primarily by the production operations. In other cases, especially in the electronics industry, inspection might take considerably longer than the actual production. An example is surface mount technology (SMT) for the assembly of printed circuit boards (see Bai and Yun [9]). An SMT circuit board is very compact and complex, consisting of hundreds or thousands of components. Compared

to actual production, inspection of these individual components for conformance is complicated. In such cases, inspection primarily determines production cycle time.

Our aim in Section 4.4.1 is to determine an inspection plan and admission policy for the general case where any of the stations can be the bottleneck. This will involve stating some assumptions about the characteristics of the production process and identifying whether or not to stabilize the individual stations in the line. Then we consider the special case where all repair stations are balanced in Section 4.4.2.

4.4.1 General Case

Consider the serial production system described in Section 4.1.1, where any of the operation, inspection, and repair stations could be a bottleneck for the system. We now show that a serial line operating under optimal conditions may indeed have unstable repair stations; however, all inspection and operation stations should be balanced. First, we state the following assumptions.

Assumption 4.4.1. *The inspection process satisfies* $\alpha_{i,j} + \beta_{i,j} \leq 1$ *for all* i *and* j.

Assumption 4.4.2. *The inspection process satisfies* $D_i \cap D_{i'} = \emptyset$ *for all stages* i, i' *with* $i \neq i'$.

Assumption 4.4.3. *For all stages* i *and units* n*, the operation, inspection, and repair station costs* $X_i^O(n)$*,* $X_i^I(n)$*, and* $X_i^R(n|z)$*, where* $z \in Z_i$*, and scrap costs* $U_i^I(n)$ *and* $U_i^R(n)$ *do not depend on the set of defects* $D \backslash D_i$ *not inspected for at stage* n*. Similarly, the revenue* $X_{N+1}^O(n)$ *for the* n*th unit does not depend on unit's defect status for all* $j \in D$ *and* n*. Finally, the limiting goodwill cost* E^G *is nondecreasing with the fraction of defective units* $\pi_{N+1,j}^O$ *for all* j.

Assumption 4.4.1 is a natural one and is needed to make sure that the inspection stations function for the benefit of the system. To see this, consider the fraction $d_{i,j}$ of items classified to be nonconforming in defect j by inspection station I_i, as given in (73). Rearranging the terms, we get $d_{i,j} = \pi_{i,j}^I(1 - \beta_{i,j} - \alpha_{i,j}) + \alpha_{i,j}$ for $j \in D_i$. If $\alpha_{i,j} + \beta_{i,j} > 1$, then as the fraction of defective units increases, the fraction classified as nonconforming decreases. As we pointed out in Section 4.2.2, various inspection policies are possible, and Assumption 4.4.1 is a refinement about the structure of these policies. Assumption 4.4.2 ensures that each defect type can be inspected for at most once. This is assumption is not very restrictive. For example, if the same defect can be introduced at different locations, then we can give the defect a different number depending on the location where it is introduced and then inspect

for the defect in multiple locations without violating Assumption 4.4.2. Also, even though Assumption 4.4.2 implies that we cannot inspect for a defect introduced at one location in several places downstream, this does not seem restrictive because if it is important to inspect for this defect, we would want to detect it as soon as possible to avoid incurring additional production costs on the defective item. Finally, Assumption 4.4.3 ensures that the cost functions depend in a sensible way on a unit's defect status. In particular, Assumption 4.4.3 states that revenue from each unit is independent of its defect structure and that goodwill cost increases as items become more defective. Assumption 4.4.3 also ensures that all costs in a given stage depend only on the set of defects inspected for at that stage (and not on other defects).

Under the above assumptions, it is beneficial to balance all operation and inspection stations in the production line, as stated in the next theorem.

Theorem 4.4.1. *Under Assumptions 4.4.1, 4.4.2, and 4.4.3, the objective function T_P (99) is maximized when all operation and inspection stations are balanced (i.e., $\lambda_i = \lambda_i^O = \lambda_i^I$ for all $i = 1, \ldots, N$).*

The proof of Theorem 4.4.1 can be found in Appendix A, and includes showing that it is better to use admission control than to allow operation or inspection stations to be unstable. However, other mechanisms, such as changing the inspection policy, may be preferable to admission control.

We note that although all operation and inspection stations should be stable, one or more of the repair stations could be unstable under the optimal conditions. This involves discarding the excess units to be repaired. To see this, consider a simple example with one operation station and associated inspection and repair stations. Minor defects are introduced at the operation station with probability p and the inspection and repair stations are error free. Also, assume that the goodwill cost is very high, so that it is optimal to inspect all units after the operation station (hence $f_1 = 1$), and that the input λ to the system yields stable operation and inspection stations, but $\lambda p > \mu_1^R$, so that the repair station is the bottleneck. When we have a push forward scheme without stabilizing the repair station, the total throughput of the system is given by $\lambda(1-p) + \mu_1^R$. However when the repair station is stabilized through admission control, the throughput reduces to $\mu_1^R(1-p)/p + \mu_1^R$. Hence, if the units are highly profitable, it is possible for the limiting profit to decrease when the repair station is stabilized.

To determine the best possible inspection points, we maximize the total profit rate function T_P (99) that combines the effects of throughput with the product quality

(our problem formulation does not take into account any fixed costs associated with including an inspection station in a particular location). In light of Theorem 4.4.1, we can adjust the flow rate calculations at each stage to stabilize the inspection and operation stations (so that $\lambda_i = \lambda_i^O = \lambda_i^I$ for all i, resulting in fewer constraints and variables). Then we can determine the flow rates, inspection locations, admission control, and desired status of repair stations (stable or unstable) using the following Nonlinear Program (NLP) whose decision variables include the inspection plan f_i, $i = 1, \ldots, N$, and the optimal amount λ_1 to admit to the system given the incoming defect information $\pi_{1,j}^O$ for all j.

$$
\begin{aligned}
\max \quad & T_P \quad \text{s.t.} \\
\lambda_1 \quad & \leq \quad \lambda; && (100)\\
\lambda_i \quad & \leq \quad \mu_i^O,\ i = 1, \ldots, N; && (101)\\
f_i\lambda_i \quad & \leq \quad \mu_i^I,\ i = 1, \ldots, N; && (102)\\
\lambda_i^R \quad & = \quad \min(\lambda_i f_i r_i^I, \mu_i^R),\ i = 1, \ldots, N; && (103)\\
\lambda_{i+1} \quad & = \quad \lambda_i^R(1 - s_i^R) + \lambda_i[1 - f_i(r_i^I + s_i^I)],\ i = 1, \ldots, N; && (104)\\
& \text{Equations } (70), (72) - (75), (81) - (87), (90) - (91),\ i = 1, \ldots, N; && (105)\\
& \text{Equations } (93) - (94), (97) - (99) \text{ for } i = 1, \ldots, N; && (106)\\
0 \quad & \leq \quad f_i \leq 1, i = 1, \ldots, N; && (107)\\
0 \quad & \leq \quad \lambda_{N+1}, \lambda_i, i = 1, \ldots, N. && (108)
\end{aligned}
$$

The objective T_P is the general total profit rate function. The constraints (100)$-$(104) represent balance equations for the flow in the serial line, allowing only the repair stations to become unstable. The constraints (105)-(106) are needed because some of the model parameters depend on the decision variables f_i, $i = 1, \ldots, N$. We also need to substitute $\lambda_I^{RO} = \lambda_i^R(1 - s_i^R)$ and $\lambda_i^{IO} = \lambda_i[1 - f_i(r_i^I + s_i^I)]$ in (91); $\lambda_i^I = \lambda_i$ and $\nu_i^I = \lambda_i f_i s_i^I$ in (93); $\nu_i^R = \lambda_i^R s_i^R$ in (94); and finally $\lambda_i^O = \lambda_i$ in (97). Let the solution be given by $\bar{f}_i, \bar{\lambda}_i$, and $\bar{\lambda}_i^R$ for $i = 1, \ldots, N$. Then the repair station R_i is stable if $\bar{\lambda}_i^R \leq \mu_i^R$ and unstable if $\bar{\lambda}_i^R > \mu_i^R$ in the allocation scenario that maximizes the return. We also reject at the rate $\nu_0^O = \lambda - \bar{\lambda}_1$ at the dummy operation node O_0, and stages i with $\bar{f}_i > 0$ are assigned an inspection station.

4.4.2 Operation or Inspection Constrained Case

In this section, we consider a serial production system as described in Section 4.1.1, under the assumption that the production capacity is determined by the operation

and inspection stations (so that all repair stations are always stable). Note that when manufacturing operations constrain the capacity of the system, the operation station O_b with the slowest production rate (so that $\mu_b^O \leq \mu_i^O$, for all i) is not necessarily the bottleneck due to the effects of inspection through scrapped units. In particular, any other operation station O_i with $i < b$ could be the bottleneck if enough units are scrapped between operation stations O_i and O_b. Similarly, the inspection station I_b with the slowest inspection rate (so that $\mu_b^I \leq \mu_i^I$, for all i) is not necessarily the bottleneck because of the effects of fractional inspection and scrapped units.

As in the general case, ensuring stability through admission control is preferred over a push forward scheme, where every operation or inspection node processes as much as possible. The next result provides the admission control that balances the operation and inspection stations for a given inspection policy $f_1, \ldots, f_N$ when repair stations never constrain the system productivity.

Theorem 4.4.2. *Suppose that $\mu_i^R \geq \lambda_{N+1} f_i r_i^I / \prod_{n=i}^N (1 - f_n s_n)$ for all i, where $s_i = s_i^I + r_i^I s_i^R$ is the total fraction of inspected units scrapped at the ith stage and*

$$\lambda_{N+1} = \min \Bigg\{ \lambda \prod_{n=1}^N (1 - f_n s_n); \mu_1^O \prod_{n=1}^N (1 - f_n s_n); \mu_2^O \prod_{n=2}^N (1 - f_n s_n); \ldots; \mu_N^O (1 - f_N s_N); \frac{\mu_1^I}{f_1} \prod_{n=1}^N (1 - f_n s_n); \ldots; \frac{\mu_N^I}{f_N} (1 - f_N s_N) \Bigg\}. \quad (109)$$

Then the line can be balanced through admission control at the dummy operation station O_0 at the rate

$$\nu_0^O = \lambda - \frac{\lambda_{N+1}}{\prod_{n=1}^N (1 - f_n s_n)}, \quad (110)$$

and λ_{N+1} is the maximum possible throughput.

Proof. When all stations are stable under the arrival rate λ, the total fraction of units scrapped at the i^{th} stage is given by $f_i s_i$, and the remaining fraction $1 - f_i s_i$ is routed to O_{i+1}. In order to achieve the throughput λ_{N+1}, we must have

$$\lambda_N = \frac{\lambda_{N+1}}{1 - f_N s_N}.$$

Continuing this argument backwards, we have at the i^{th} stage

$$\lambda_i = \frac{\lambda_{N+1}}{\prod_{n=i}^N (1 - f_n s_n)}.$$

Then by equation (79), the condition on μ_i^R ensures the stability of all repair stations. Each operation node is constrained by its production capacity, and the dummy operation node O_0 is constrained by the arrival rate λ, so stability requires

$$\begin{aligned}
\frac{\lambda_{N+1}}{\prod_{n=i}^{N}(1-f_n s_n)} &\leq \mu_i^O,\ i=1,\ldots,N;\\
\frac{\lambda_{N+1}}{\prod_{n=i}^{N}(1-f_n s_n)} &\leq \frac{\mu_i^I}{f_i},\ i=1,\ldots,N;\\
\frac{\lambda_{N+1}}{\prod_{n=1}^{N}(1-f_n s_n)} &\leq \lambda.
\end{aligned}$$

Hence the equalities (109) and (110) follow. □

Finally, based on the previous two theorems and the incoming defect information $\pi_{1,j}^O$ for all j, we construct the following NLP for finding the best inspection policy when all repair stations are known to be stable. This would be the case, for instance, when all defects are considered major.

$$\begin{aligned}
\max \quad & \lambda_{N+1}R - \lambda_{N+1}\sum_{w\in\{0,1\}^{|D|}}\tilde{P}(W_{N+1}=w)E_G(w)\\
& -\sum_{i=1}^{N}\frac{\lambda_{N+1}}{\prod_{n=i}^{N}(1-f_n s_n)}\left[r_i^I f_i\left(\sum_{z\in Z_i}\tilde{P}(z)E_i^R(z)+s_i^R U_i^R\right)+E_i^O+f_i E_i^I+f_i s_i^I U_i^I\right] \qquad (111)
\end{aligned}$$

where

$$\begin{aligned}
\lambda_{N+1} &\leq \lambda\prod_{n=1}^{N}(1-f_n s_n); && (112)\\
\lambda_{N+1} &\leq \mu_i^O\prod_{n=i}^{N}(1-f_n s_n),\ i=1,\ldots,N; && (113)\\
\lambda_{N+1} &\leq \frac{\mu_i^I}{f_i}\prod_{n=i}^{N}(1-f_n s_n),\ i=1,\ldots,N; && (114)\\
&\text{the set of constraints (105);} && (115)\\
s_i &= s_i^I + r_i^I s_i^R \text{ for } i=1,\ldots,N; && (116)\\
0 &\leq f_i \leq 1,\ i=1,\ldots,N; && (117)\\
0 &\leq \lambda_{N+1}. && (118)
\end{aligned}$$

Constraint (112) means that the throughput for the system is limited by the total available demand λ. Similarly, constraints (113) and (114) represent the capacity

limitations at each of the operation and inspection stations, respectively. Let the solution be given by $\bar{f}_i$ and $\bar{\lambda}_{N+1}$. If the system is profitable, so that $\bar{\lambda}_{N+1} > 0$, then $\bar{\lambda}_{N+1}$ can be obtained as in Theorem 4.4.2 since the objective function in (111) is linear in λ_{N+1} given f_i for all i; otherwise we have $\bar{\lambda}_{N+1} = 0$. Note that this NP is simpler than the general problem in (100) − (108), with fewer variables and constraints, as a result of the assumption that repair stations are never unstable.

4.5 A Numerical Example

In this section, our aim is to gain insights into the behavior of the optimal inspection allocation strategy, as well as to demonstrate the effects of throughput considerations on the optimal inspection allocation. We consider a basic model with $N = 2$ operation stations O_1 and O_2 in tandem and two possible inspection locations after the operation stations. We have two types of defects that can be introduced independent of each other in the production process with $p_{1,1} = p_1$, $p_{1,2} = 0$ and $p_{2,1} = 0$, $p_{2,2} = p_2$. The inspection process is assumed to be all or none, i.e., $f_1, f_2 \in \{0, 1\}$. For simplicity we assume that defective units are detected during inspection without any error (i.e., $\alpha_{i,j} = \beta_{i,j} = 0$, $\forall i, j$). Moreover, both defect types are assumed to be major, so that all defective units are scrapped without any repair attempt, and unit inspection costs are additive. Our aim is to find an inspection allocation policy with admission control that maximizes the profit rate of the system. Then the possible actions are

$$
\begin{aligned}
a_0 &= \text{No inspection } (f_1 = f_2 = 0);\\
a_1 &= \text{Inspect for defect 1 after operation station 1 } (f_1 = 1, f_2 = 0);\\
a_2 &= \text{Inspect for defect 2 after operation station 2 } (f_1 = 0, f_2 = 1);\\
a_3 &= \text{Inspect for defect 1 after operation station 1 and for defect 2}\\
&\quad\ \text{after operation station 2 } (f_1 = 1, f_2 = 1).
\end{aligned}
$$

Note that, we do not consider inspecting for defect 1 only after operation station 2 since action a_1 will always outperform this case in our model. Similarly, action a_3 is always better than inspecting for both defects 1 and 2 after the operation station 2 (since we assume that inspection costs are additive and that there are no fixed costs associated with inspecting in two locations). As in the general model, each unit generates a revenue R, as well as goodwill costs G_1 and G_2 depending on whether it has defect 1 or defect 2, respectively. If a unit has both defects, we assume a goodwill cost of $G_1 + G_2$ is introduced. Processing costs at operation stations 1 and 2 include any cost associated with production and are represented by C_1 and C_2, respectively.

Inspecting at location 1 (2) incurs an inspection cost of H_1 (H_2) per unit. Each scrapped unit incurs a scrap cost given by U_1 or U_2 depending on whether it was scrapped at the first or second inspection location. Since the scrap cost U_2 includes both U_1 and the processing cost at operation station 2, we assume that $U_2 \geq U_1$.

Given the above assumptions, we can formulate the profit rate function T_P depending on the inspection strategy as well as the input parameters, R, C_i, G_i, H_i, and U_i, $i = 1, 2$. Let $\lambda_0 = \lambda$ denote the arrival rate to the system and $\mu_1^O = \mu_1$, $\mu_2^O = \mu_2$ be the processing capacities at stations 1 and 2, respectively. To observe the effects of inspection operations on a system constrained by the capacity of a bottleneck operation station, we assume that μ_1^I, $\mu_2^I \geq \lambda_0$, so that the inspection operations never constrain the system. The throughput λ_{N+1} of the system is denoted by μ.

The optimal inspection locations are determined based on the NLP $(111) - (118)$, except f_i, $i = 1, 2$, are now constrained to be integers. Since all defects are major and inspection is error free, we have $s_i^I = s_i = p_i$, and also $r_i^I = s_i^R = 0$ for $i = 1, 2$. Moreover, $E_i^O = C_i$, $E_i^I = H_i$, $U_i^I = U_i$, $\pi_{1,1}^O = 0$, $\pi_{1,1}^I = p_1$, $\pi_{1,2}^O = \pi_{1,2}^I = 0$, $\pi_{2,1}^O = \pi_{2,1}^I = p_1(1 - f_1)$, $\pi_{2,2}^O = 0$, $\pi_{2,2}^I = p_2$, and $\pi_{3,j}^I = \pi_{3,j}^O = p_j(1 - f_j)$ for $i = 1, 2$ and $j = 1, 2$. Hence NLP $(111) - (118)$ becomes

$$\begin{aligned}
\max \quad & \mu R - \sum_{i=1}^{2} \frac{\mu f_i}{\prod_{n=i}^{2}(1 - f_n s_n)} \left(\frac{C_i}{f_i} + H_i + p_i U_i \right) \\
& - \mu \Big(G_1 p_1 (1 - f_1) + G_2 p_2 (1 - f_2) \Big) \qquad (119) \\
& \text{such that} \\
& \mu \leq \lambda (1 - f_1 p_1)(1 - f_2 p_2), \qquad (120) \\
& \mu \leq \mu_1 (1 - f_1 p_1)(1 - f_2 p_2), \qquad (121) \\
& \mu \leq \mu_2 (1 - f_2 p_2), \qquad (122) \\
& f_1, f_2 \in \{0, 1\}, \text{ and } \mu \geq 0. \qquad (123)
\end{aligned}$$

Note that in the above allocation NLP, all variables are known except for the decision variables f_1, f_2 and the throughput μ. Goodwill cost in the objective function (119) results from the fact that goodwill costs are additive and we have no inspection errors. Since the problem size is small, we can easily solve the NLP in $(119) - (123)$ by enumerating all four possible solutions and obtain the profit function $T_P(f_1, f_2)$.

In particular,

$$
\begin{aligned}
T_P(0,0) &= \min\{\lambda,\mu_1,\mu_2\}(R - C_1 - C_2 - G_1p_1 - G_2p_2), && (124)\\
T_P(1,0) &= \min\{(1-p_1)\lambda,(1-p_1)\mu_1,\mu_2\}\left(R - \frac{C_1+H_1+p_1U_1}{1-p_1} - C_2 - G_2p_2\right), && (125)\\
T_P(0,1) &= (1-p_2)\min\{\lambda,\mu_1,\mu_2\}\left(R - \frac{C_1}{1-p_2} - \frac{C_2+H_2+p_2U_2}{1-p_2} - G_1p_1\right), && (126)\\
T_P(1,1) &= (1-p_2)\min\{(1-p_1)\lambda,(1-p_1)\mu_1,\mu_2\}\\
&\quad\times\left(R - \frac{C_1+H_1+p_1U_1}{(1-p_2)(1-p_1)} - \frac{C_2+H_2+p_2U_2}{1-p_2}\right). && (127)
\end{aligned}
$$

However, since these functions involve many variables, it is not trivial to obtain general structural results. Instead, we will visualize the optimal actions as functions of p_1 and p_2 given the input parameters λ, μ_i, R, G_i, H_i, and U_i, $i = 1, 2$.

We will consider two cases. In the first case, production is constrained by the arrival rate λ to the system, and in the second case, the second operation station is a bottleneck and determines the throughput. In this case, as shown in Section 4.4, we control the input rate to the system to balance the operation stations (note that the rejection of arriving units does not incur any penalty costs in our model). We do not consider the case where the first station is the bottleneck because this case is the same as the first case except that the arrival rate is now controlled to equal the processing capacity of the first station.

Next, we provide the total profit functions in input and capacity constrained systems. For the first case where the system is constrained by the arrival rate, equations $(124) - (127)$ become

$$
\begin{aligned}
T_P(0,0) &= \lambda(R - C_1 - C_2 - G_1p_1 - G_2p_2),\\
T_P(1,0) &= \lambda(1-p_1)\left(R - \frac{C_1+H_1+p_1U_1}{1-p_1} - C_2 - G_2p_2\right),\\
T_P(0,1) &= \lambda(1-p_2)\left(R - \frac{C_1}{1-p_2} - \frac{C_2+H_2+p_2U_2}{1-p_2} - G_1p_1\right),\\
T_P(1,1) &= \lambda(1-p_2)(1-p_1)\left(R - \frac{C_1+H_1+p_1U_1}{(1-p_2)(1-p_1)} - \frac{C_2+H_2+p_2U_2}{1-p_2}\right).
\end{aligned}
$$

Note that the actual value of λ and the capacity of the operation stations have no effect on the optimal decision, since we can always compare the functions $T_P(\cdot)/\lambda$. In this case, we cannot see the side benefit of having inspection before the bottleneck on the optimal decision and our inspection allocation decision agrees with traditional wisdom in that we choose to inspect whenever the expected inspection cost for a unit is less than the expected cost of not inspecting (see, e.g., Bai and Yun [9], Chen

[22], Eppen and Hurst [56], Lindsay and Bishop [66], and Raz and Kaspi [81]). More specifically, rearranging the profit functions, we obtain

$$\begin{aligned}
\frac{T_P(1,0)-T_P(0,0)}{\lambda} &= p_1(C_2+G_2p_2+G_1)-(H_1+p_1R+p_1U_1), \qquad (128)\\
\frac{T_P(0,1)-T_P(0,0)}{\lambda} &= p_2(G_1p_1+G_2)-(H_2+p_2R+p_2U_2),\\
\frac{T_P(1,1)-T_P(0,0)}{\lambda} &= (G_1p_1+G_2p_2+C_2p_1)\\
&\quad -(H_1+p_1U_1+(H_2+p_2U_2)(1-p_1)+R[p_1+p_2-p_1p_2]),
\end{aligned}$$

where $(H_i+p_iR+p_iU_i)$, $i=1,2$, can be considered as the expected cost of inspection for defect i only per unit, and $(H_1+p_1U_1+(H_2+p_2U_2)(1-p_1)+R[p_1+p_2-p_1p_2])$ as the expected cost of inspection when inspecting both defects simultaneously. Similarly, $p_1(C_2+G_2p_2+G_1)$, $p_2(G_1p_1+G_2)$, $(G_1p_1+G_2p_2+C_2p_1)$ can be viewed as the expected cost of not inspecting at locations 1, 2 and, 1 and 2 together, respectively.

Next we derive the profit functions for the interesting case where the production line is constrained by the capacity of the second operation station having processing rate μ_2, assuming that $\mu_1>\lambda>\mu_2/(1-p_1)$. The first condition ensures that O_1 is not a bottleneck, and the second condition ensures that we have in all cases enough input for O_2 to be a bottleneck. Hence, as a result of Theorem 4.4.1, we reject any incoming job with probability $(\lambda-\mu_2/(1-p_1f_1))/\lambda$ to stabilize the system. In this case, we can observe the effects of throughput considerations on the inspection allocation decision. In particular, the profit functions are given by

$$\begin{aligned}
T_P(0,0) &= \mu_2(R-C_1-C_2-G_1p_1-G_2p_2), \qquad (129)\\
T_P(1,0) &= \mu_2\left(R-\frac{C_1+H_1+p_1U_1}{1-p_1}-C_2-G_2p_2\right), \qquad (130)\\
T_P(0,1) &= \mu_2(1-p_2)\left(R-\frac{C_1}{1-p_2}-\frac{C_2+H_2+p_2U_2}{1-p_2}-G_1p_1\right),\\
T_P(1,1) &= \mu_2(1-p_2)\left(R-\frac{C_1+H_1+p_1U_1}{(1-p_2)(1-p_1)}-\frac{C_2+H_2+p_2U_2}{1-p_2}\right).
\end{aligned}$$

As before, the actual value of μ_2 has no effect on the optimal decisions as long as O_2 remains a bottleneck, because we can always compare the functions $T_P(\cdot)/\mu_2$. Note that, by inspecting for defect one after the first operation station and scrapping defective items, we not only remove the defective items but also increase the capacity of the production line from μ_2 to $\mu_2/(1-p_1)$. Thus, in the capacity constrained case, inspecting after the first operation has gained an advantage with magnitude depending on the value p_1. In all prior works on inspection allocation, this secondary

effect is neglected, although the production line might be constrained by one of the operation stations and not the external input. Because of this secondary effect, we will see that even under the same defect distribution and cost parameters, inspecting after operation station 1 for defect 1 becomes more beneficial as compared to the first case.

Goodwill cost is a good measure of how important certain quality characteristics are to customers. Our experience with solving the NLP $(119)-(123)$ reveals that the relative values of G_1, G_2, and R are important factors in inspection allocation decisions. Hence, we compare the inspection allocation decisions in capacity constrained and input constrained systems with different relative values for G_1, G_2, and R, namely when there is at least one serious defect (i.e., $\max\{G_1, G_2\} > R$) and when both defect types are considered not serious (i.e., $\max\{G_1, G_2\} < R$), in Figures 10 and 11, respectively. The specific values for G_1, G_2, and R are chosen so that a number of different optimum actions are observed. Each shaded region in Figures 10 and 11 is labeled with the optimal actions for the corresponding defect probabilities p_1 (shown on the x-axis) and p_2 (shown on the y-axis); unshaded regions correspond to the case when the system is not profitable under any of the actions $a_0, \ldots, a_3$ (so that $\mu = 0$). Note that the possibility of having at least one serious defect such that the associated goodwill cost is higher than the revenue is possible, for instance, when the defect causes the company not only to lose the revenue but also to incur some additional costs, like repair and shipping costs or loss of company reputation. In both figures, the left column shows the optimal actions when the system is input constrained and the right column depicts the optimal actions when the second station is a bottleneck. Note that the arrival rate λ and processing capacities μ_1, μ_2 do not affect the optimal decisions beyond determining whether the system is arrival constrained or constrained by the capacity of the second station. Throughout we choose $\mu_1 = 110$, $\mu_2 = 1$, $H_1 = H_2 = 2, U_1 = U_2 = 4$, and $C_1 = C_2 = 5$. Moreover, $\lambda = 1$ for input constrained systems, and $\lambda = 100$ when the second station is a bottleneck.

In all cases, whether input or capacity constrained, we observe that there is no inspection when the defect probabilities p_1 and p_2 are low. However, as defect probabilities associated with high goodwill costs increase, we choose to add inspection stations.

In Figure 10, we study cases where at least one of the defects is serious. In particular, we consider $G_2 < R < G_1$ in parts (a)-(b) of the figure, $G_1 < R < G_2$ in parts (c)-(d), and $R < G_1 < G_2$ in parts (e)-(f); the case $R < G_2 < G_1$ produces a graph that is very similar to the one in Figure 10 (e)-(f), and hence is not included.

We note that even in this case, inspecting for both defects is not the best option for the demand constrained system, unless G_1 and G_2 are both high compared to R and the defect probabilities p_1 and p_2 are large. This does not necessarily hold for the capacity constrained system, as shown in Figure 10 (d), where $G_1 < R < G_2$ and inspecting for both defects is desirable for some values of p_1, p_2. The fact that actions a_1 and a_3 are in some instances optimal in the capacity constrained case, but not in the corresponding demand constrained case, even when G_1 is not high, is intuitive because inspecting for defect 1 has the additional benefit of increasing system capacity. Existence of a serious defect implies inspection for that defect unless the associated defect fraction is low, see Figure 10 (a)-(b), (c)-(d), and (e)-(f), where defect 1, defect 2, and defects 1 and 2 are serious, respectively. Thus, when both defect types are serious, as in Figure 10 (e)-(f), then inspection at both locations is required when both defect fractions are large. Also, note that the region where inspecting for both defects is best is never adjacent to the no inspection region, see parts (d)-(f) of Figure 10, which means that as the defect probabilities change, we never jump from no inspection to inspection for both defects (there is always an intermediate step, where we inspect for only one of the defects).

In Figure 11, we consider cases where both of the defect types are not serious, i.e., when $G_2 < G_1 < R$ (parts (a) and (b)) and when $G_1 < G_2 < R$ (parts (c) and (d)). We note that when the goodwill costs are low, the decision not to inspect is often optimal, even if the inspection costs are low, because we do not want to scrap units without any revenue. As a result, no inspection is preferred for all values of the defect probabilities in Figure 11 (a) and for most values of the defect probabilities in Figure 11 (c) when the system is input constrained. Note that we choose to inspect for defect 1 for some values of defect probabilities in Figure 11 (c), but not in Figure 11 (a), even though the goodwill cost G_1 is higher in case (a). This behavior is explained by the fact that the choice to inspect at location 1 is also affected by the value of goodwill cost 2, see (128). However, when the system is capacity constrained, even though goodwill costs are low, inspection decision after O_1 is favored, see Figure 11 (b) and (d). This is a result of increasing production capacity by scrapping or repairing defective items before a bottleneck operation station, and hence reducing the waste of operation capacity on defective products. Finally, inspecting for both defects is not preferred in all cases since both defects are not serious.

In both Figures 10 and 11, we can clearly see the effects of throughput consideration when allocating the inspection stations. In particular, since the throughput increases only if we inspect after the first operation station, actions a_1 and a_3 are

optimal for larger ranges of defect probabilities in the capacity constrained case. In Figure 10, part (d), we observe that it is more beneficial to inspect after operation station 1 when the defect probability p_1 is not close to one of the extreme points 0 or 1. This is due to the fact that while the gain associated with inspecting after O_1 is linear in p_1, the loss grows exponentially as p_1 increases in the capacity constrained case. In particular, when $\mu_1 > \lambda > \mu_2/(1-p_1)$, we have

$$\frac{T_P(1,0) - T_P(0,0)}{\mu_2} = p_1 G_1 - \frac{C_1 p_1 + H_1 + U_1 p_1}{1-p_1}, \tag{131}$$

see (129) − (130). The value of p_1 where the difference in (131) is maximized is $p_1^* = 1 - \sqrt{(C_1 + H_1 + U_1)/G_1}$. For the example in Figure 10, part (d), this point is given by $p_1^* = 0.475$. Hence, contrary to what might be expected, action a_1 is most beneficial when the defect probability p_1 is within a certain range, and as p_1 gets closer to 1, action a_1 ceases to be optimal. Similar behavior is observed in Figure 11 (b) and (d), where a_1 is not optimal for high and low values of p_1.

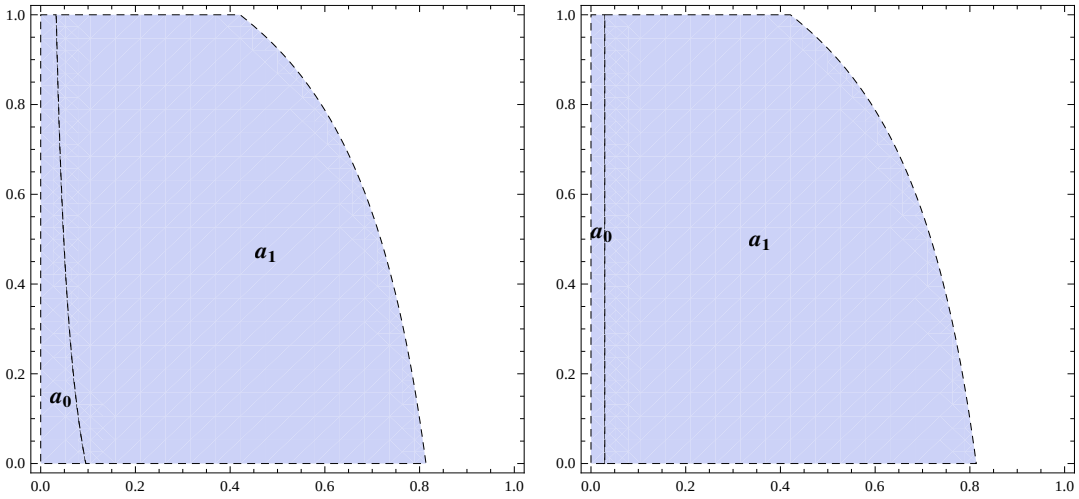
Finally, parts (a) and (b) of Figure 12 show the difference in profit between the capacity constrained and input constrained cases under optimal inspection decisions as functions of the defect fractions p_1 and p_2 when defect 2 is the only serious defect (as in parts (c) and (d) of Figure 10) and when there are no serious defects (as in parts (c) and (d) of Figure 11), respectively. In both cases, darker regions correspond to higher difference values. For instance, in Figure 5 (a), when $p_1 = 0.45$ and $p_2 = 0.02$, the total profits are 30.4 and 37.4 in the input (no inspection) and capacity constrained (inspect at location 1) cases, respectively. Thus, we observe an increase of 23% in profit when we take the capacity of the system into account in making inspection decisions. Similarly, when $p_1 = 0.5$ and $p_2 = 0.4$, the total profits are 36.5 and 46 in the input (no inspection) and capacity constrained (inspect at location 1) cases, respectively, corresponding to a increase of 26% in profit. Thus, taking capacity considerations into account while making inspection allocation decisions can have a substantial impact on profit. The increase in total profit is attributed to the fact that inspection before a bottleneck station can improve the throughput of the system.

4.6 Conclusion

Product quality is a vital consideration for any manufacturing firm aiming to keep a competitive edge. However, maintaining high product quality can be expensive. As a result, effective inspection location choices have traditionally depended on the tradeoff between inspection costs and goodwill costs incurred as a result of poor

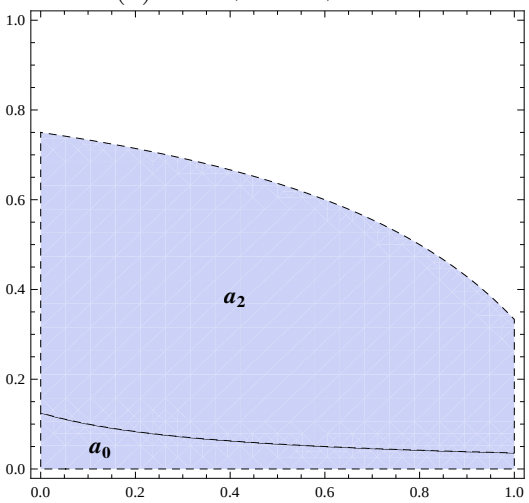
product quality. However, this ignores the secondary effects of inspection on the production capacity of a system. By contrast, our analysis accounts for the effects of inspection on both quality and quantity simultaneously. More specifically, we showed how to calculate product flows and fraction of defective units at each production stage in a step-by-step manner. Using the flow and defect information, we computed the various costs incurred throughout the serial line, as well as the resulting profits. Moreover, our model is more general than any model considered previously in the literature, including multiple defect types, defect classifications (major and minor), defect-dependant inspection errors, fractional inspection, probabilistic repairs that are defect dependent, and stochastic operation, inspection, repair, and goodwill costs, as well as revenue.

Under mild assumptions, we showed that under the optimal inspection policy, operation and inspection stations should be stable, while repair stations can be unstable. We also formulated nonlinear programs for determining the optimal inspection allocation and admission policies for both the general case and when all stations are stable. Finally, through numerical examples, we studied the effects of taking capacity into account when choosing an inspection policy. Our numerical results show that ignoring the effects of inspection on capacity can result in suboptimal inspection location decisions, leading to substantial decreases in profit.

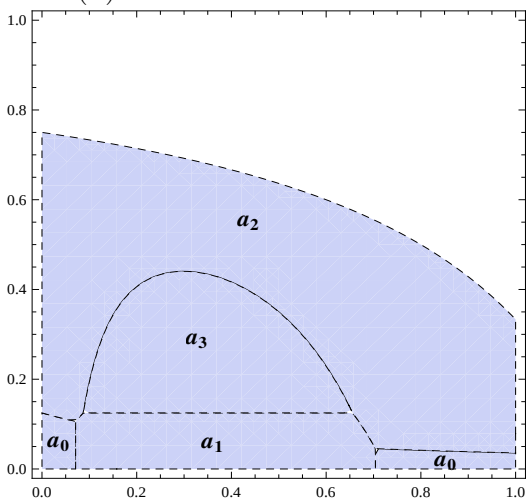


(a) $R = 60, G_1 = 80, G_2 = 40$

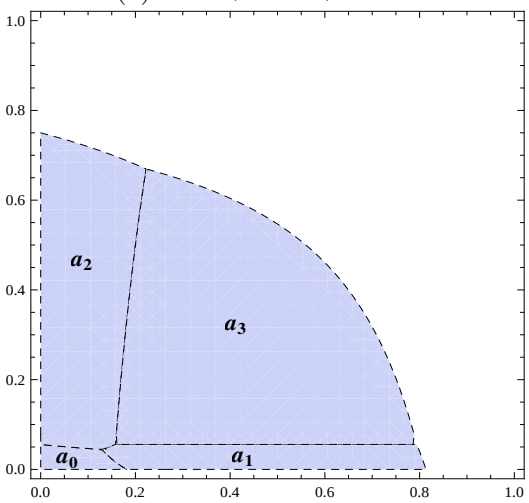
(b) When Station 2 is a Bottleneck



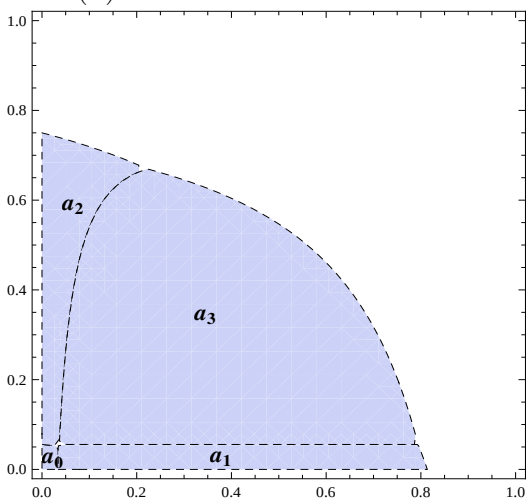
(c) $R = 60, G_1 = 40, G_2 = 80$



(d) When Station 2 is a Bottleneck

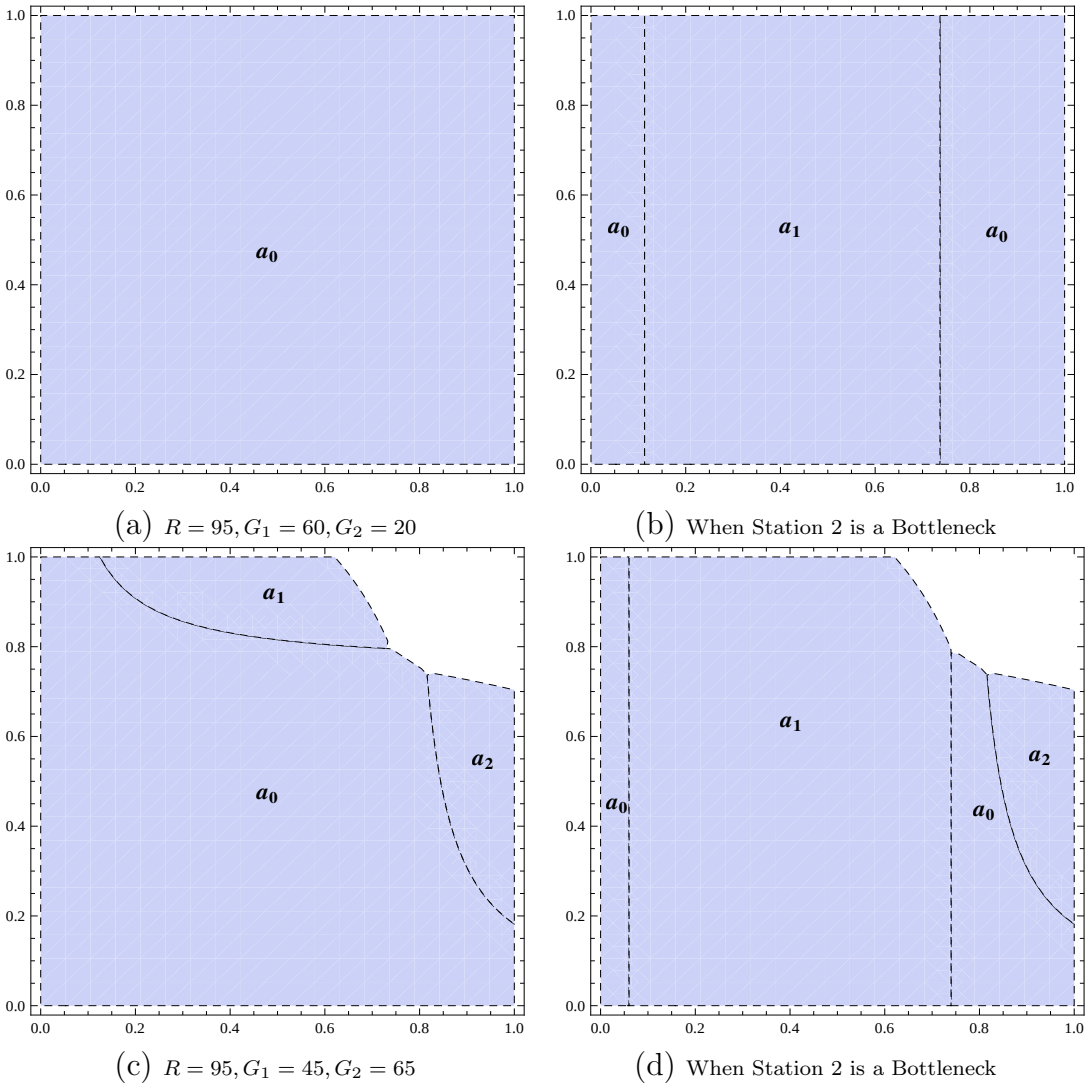


(e) $R = 60, G_1 = 70, G_2 = 100$



(f) When Station 2 is a Bottleneck

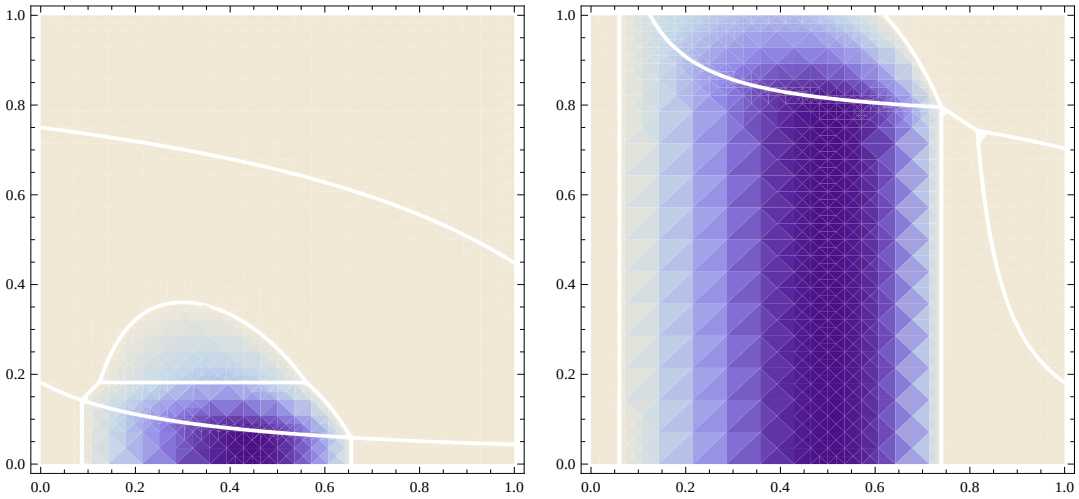
Figure 10: Comparison of optimal inspection allocation strategies as a function of p_1 and p_2 for systems with at least one serious defect



(a) $R = 95, G_1 = 60, G_2 = 20$ (b) When Station 2 is a Bottleneck

(c) $R = 95, G_1 = 45, G_2 = 65$ (d) When Station 2 is a Bottleneck

Figure 11: Comparison of optimal inspection allocation strategies as a function of p_1 and p_2 for systems without a serious defect



(a) $R = 60, G_1 = 40, G_2 = 80$ (b) $R = 95, G_1 = 45, G_2 = 65$

Figure 12: Magnitude of the difference for the profit functions in input and capacity constrained cases

CHAPTER V

CAPACITY SIZING AND PRICING WITH HETEROGENOUS PRODUCTS AND RESOURCES

In today's highly competitive market, the capacity investment decision is an important factor affecting a firm's profitability and competitiveness. At the time of the capacity decision, uncertainty in demand results from consumer preferences not being observable and uncertainty in economic conditions (see Bish, Liu, and Suwandechochai [16]). Firms are increasingly resorting to flexibility, on both the supply and demand sides, to effectively match their supply with demand. In this chapter, we study the capacity sizing problem faced by a price setting and monopolistic firm producing two substitutable/complementary products with flexible and dedicated technologies.

The problem of determining the optimal capacity under demand uncertainty has received significant attention in the literature under various assumptions (see Section 2.3). However, past research assumes that products and resources are homogenous in that all processing rates are equal. Thus, the effects of differences in flexible and dedicated resources' service rates at the product groups have been ignored. The assumption that all processing rates are equal is reasonable when resources are considered as inventories that supply all product groups indifferently, but it is restrictive when the resources have varying capabilities at the two product groups. Our aim is to explicitly model the different rates at which flexible and dedicated resources can supply the product classes, along with product substitutability and demand correlation, and see the effects of substitutability and correlation on the optimal capacity, allocation decisions, prices, and the corresponding expected profits.

More specifically, consider a firm selling two products that needs to determine the amount of production capacity to acquire at a time when little information on product demand is available. The capacity decision made at this first strategic stage constrains the firm's capabilities when demand information becomes available. We use a linear demand model, where the demand for each product is inversely related to its own price, with possible cross-price effects from the other product. Demand uncertainty is modeled as uncertainty about the location of demand lines. This problem can be formulated as a two-stage stochastic program. In the first stage,

before the uncertainty is resolved, the firm needs to determine the amount of dedicated capacity for products 1 and 2, as well as the flexible capacity, so as to maximize its expected profit. Then, in the second stage after the demands are observed, the production quantities, the corresponding prices, and the allocation of the servers are determined to maximize the revenue. Note, however, that the prices may be determined by market conditions, and hence it is not always possible to alter prices to change demand.

The outline of this chapter is as follows. In Section 5.1, we provide our problem formulation, including the model description and assumptions. In Section 5.2, we classify the optimal actions for the general case of our problem with dedicated and flexible servers and substitutable products. Then, we look at the special case where there is only a finite set of possibilities for the random demand intercept in Section 5.3. In Section 5.4, we present our numerical results. Finally, we summarize our findings in Section 5.5 and provide proofs of most of our results in Appendix B.

5.1 Problem Formulation

Consider a firm selling two products that needs to determine the amount of production capacity to acquire at a time when little information on product demand is available. The capacity decision made at this first strategic stage constrains the firm's capabilities when demand information becomes available. This problem can be formulated as a two-stage stochastic program. In the first stage, before the uncertainty is resolved, the firm needs to determine the amount of dedicated capacity n_1 and n_2 for products 1 and 2, as well as the flexible capacity n_f, with unit costs c_1, c_2, and c_f, respectively, so as to maximize its expected profit $V(\boldsymbol{n})$, where $\boldsymbol{n} = (n_1, n_2, n_f)$. For simplicity, the capacities $\boldsymbol{n}$ are allowed to be fractional and not restricted to be integers. At the time of the capacity investment decision, uncertainty in demand results from consumer preferences not being observable and uncertainty in economic conditions (see Bish, Liu, and Suwandechochai [16]). Expected profit is the expected optimum revenue minus the capacity investment costs. We assume that $c_1, c_2 < c_f$ and $c_f < c_1 + c_2$, because otherwise some resource type is automatically ignored in the optimal solution. As in the earlier literature (see Bish and Wang [18], Chod and Rudi [28], Fine and Freund [38], and Van Mieghem and Dada [92]), we assume a linear form for the cost of capacity acquisitions. This is without loss of generality as long as the cost of capacity acquisitions can be represented by a convex function, so that the concavity of the objective function is preserved and the KKT conditions

hold (see the proof of Theorem 5.3.1). Then, in the second stage after the demand uncertainty is resolved, the production quantities Q_1, Q_2, the corresponding prices P_1, P_2, and the allocations of the servers are determined to maximize the revenue, which is a function of $\boldsymbol{n}$ and the random demand intercepts.

Each product can be manufactured by both dedicated and flexible resources. Servers have different capabilities at the two product groups. A server dedicated to product $i \in \{1, 2\}$ can work at rate μ_i per production period. We assume, without loss of generality, that $\mu_1 \geq \mu_2$. Similarly, the service rate for a flexible server is $f\mu_i$ for product $i \in \{1, 2\}$, where $f > 0$. In practice, flexible servers are usually slower than the corresponding dedicated ones, hence we will be primarily interested in the case where $f \leq 1$. Thus our model allows both the products and the servers to be heterogenous, a major extension over prior works that only consider the case where $\mu_1 = \mu_2$ and $f = 1$.

We assume that the uncertain demand for each product can be represented as a linear function of its own price and the price of the other product, with known slopes but random y-intercepts. That is, the demand D_i for product $i = 1, 2$, is given by

$$D_1 = \xi_1 - \alpha_1 P_1 + \beta P_2, \tag{132}$$

$$D_2 = \xi_2 - \alpha_2 P_2 + \beta P_1, \tag{133}$$

where P_i is the price for product i, $\alpha_i > 0$ and β are the known own-price and cross-price elasticity parameters, respectively, and $\xi_i \geq 0$ is the random demand intercept, or the potential market size for product i when both prices are zero. We allow the products to have different own-price elasticities α_1 and α_2, but as in the related literature (see, e.g., Birge, Drogosz, and Duenyas [14], Bish and Suwandechochai [17], and Chod and Rudi [28]), the cross-price effects between the two products are symmetric, modelled by the parameter β. The cross-price elasticity β takes into account the substitutability and complementary effects across products. A positive β indicates that two products are substitutes, while negative β indicates that two products are complements. Throughout the chapter we focus on the case with $\beta \geq 0$. The case with $\beta < 0$ is covered in the numerical results section. Since the effects of a product's own price on its demand should be more than the effects of the other product's price, we assume $\alpha_i > |\beta|$ for $i = 1, 2$.

The demand uncertainty is included in the model through the random demand intercepts ξ_1 and ξ_2, where ξ_i is a non-negative random variable with mean m_i and standard deviation σ_i for $i \in \{1, 2\}$. Note that the demand intercepts ξ_1 and ξ_2 for the products might be correlated with correlation coefficient ρ. Without loss of generality,

we do not include unit production costs, because they can always be incorporated by modifying the demand intercepts $\boldsymbol{\xi}$ (i.e., if k_1 and k_2 are unit production costs for products 1 and 2, respectively, then we would use the modified demand intercepts $\xi_1' = \xi_1 + (\beta k_2 - \alpha_1 k_1)$ and $\xi_2' = \xi_2 + (\beta k_1 - \alpha_2 k_2)$. We let $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2)'$ represent a realization for (ξ_1, ξ_2), where $'$ denotes the transpose of a matrix. This demand model is commonly used in the literature (see, e.g., Bish and Wang [18], Chod and Rudi [28], Fine and Freund [38], Goyal and Netessine [44, 45], and Van Mieghem and Dada [92]).

Responsive production and pricing ability for the firm implies that prices can be modified after the demand curve intercepts $\boldsymbol{\xi} = (\xi_1, \xi_2)'$ are realized. Hence, under the assumptions (132) – (133), it is always best for the firm to match supply and demand (so that $Q_i = D_i$ for $i = 1, 2$) to maximize revenue for substitutable products. Similarly, for complementary products, we assume that all the demand is satisfied. Given the output vector $\boldsymbol{Q} = (Q_1, Q_2)'$, the corresponding prices $\boldsymbol{P} = (P_1, P_2)'$ can be determined from (132) – (133) as

$$\boldsymbol{P} = \boldsymbol{H}[\boldsymbol{\xi} - \boldsymbol{Q}], \text{ where } \boldsymbol{H} = \frac{1}{d}\begin{pmatrix} \alpha_2 & \beta \\ \beta & \alpha_1 \end{pmatrix} \text{ and } d = \alpha_1\alpha_2 - \beta^2. \tag{134}$$

Note that $\boldsymbol{H}$ is symmetric with diagonal entries that are positive and larger than the absolute value of the off-diagonal entries (because $\alpha_1, \alpha_2 > |\beta|$). It follows that $\boldsymbol{H}$ is positive definite.

Let $\boldsymbol{x} = (y_1, y_2, z_1, z_2)$ denote the product quantity vector in stage 2, where y_i and z_i represent the amount of product $i \in \{1, 2\}$ produced by dedicated and flexible resources, respectively. Also let δ_i denote the fraction of a flexible server's time devoted to product $i \in \{1, 2\}$. Given the capacities $\boldsymbol{n}$, demand intercepts $\boldsymbol{\xi}$, and $\boldsymbol{\delta} = (\delta_1, \delta_2)'$, the optimal price and output quantities can be determined in stage 2 by

$$R^*(\boldsymbol{n}, \boldsymbol{\varepsilon}) = \max_{\boldsymbol{x}, \boldsymbol{\delta}, \boldsymbol{P}} \; R(\boldsymbol{n}, \boldsymbol{\varepsilon}) = \begin{pmatrix} y_1 + z_1 \\ y_2 + z_2 \end{pmatrix}' \boldsymbol{P} \tag{135}$$

s.t.

$$\begin{aligned} y_i &\leq n_i\mu_i, i = 1, 2; && (136)\\ z_i &\leq f n_f \delta_i \mu_i, i = 1, 2; && (137)\\ \delta_1 + \delta_2 &\leq 1; && (138)\\ y_i + z_i &= \varepsilon_i - \alpha_i P_i + \beta P_j, i = 1, 2; j = 3 - i; && (139)\\ y_i, z_i, P_i, \delta_i &\geq 0, i = 1, 2. && (140) \end{aligned}$$

In the above formulation, constraints (136), (137), and (138) result from capacity

limitations. The constraints in (139) imply that the production quantity should be equal to the total demand. Finally, the constraints in (140) are the nonnegativity constraints for quantities, prices, and allocations. Note (139) and (140) imply that demand cannot be negative. Moreover, quantities produced and sold are not restricted to be integer.

Let $n_{ie} = n_i + fn_f$ and $n_e = n_1 + n_2 + fn_f$ be the total effective capacity for product $i \in \{1, 2\}$ when all flexible capacity is assigned to product i and total effective capacity, respectively. Noting that $Q_i = y_i + z_i$, we can construct the following equivalent formulation, where the decision variables are the production quantities, Q_i, $i = 1, 2$.

$$R^*(\boldsymbol{n}, \boldsymbol{\varepsilon}) = \max_{\boldsymbol{Q}} \; R(\boldsymbol{Q}) = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \boldsymbol{H} \begin{pmatrix} \varepsilon_1 - Q_1 \\ \varepsilon_2 - Q_2 \end{pmatrix} \tag{141}$$

s.t.

$$Q_1 \leq \mu_1 n_{1e}; \tag{142}$$

$$Q_2 \leq \mu_2 n_{2e}; \tag{143}$$

$$\frac{Q_1}{\mu_1} + \frac{Q_2}{\mu_2} \leq n_e; \tag{144}$$

$$\boldsymbol{H}[\boldsymbol{\varepsilon} - \boldsymbol{Q}] \geq 0; \tag{145}$$

$$Q_1, Q_2 \geq 0. \tag{146}$$

Then in stage 1, the capacity sizing problem can be formulated as

$$\max_{\boldsymbol{n}} \quad E[R^*(\boldsymbol{n}, \boldsymbol{\xi})] - \sum_{i=1,2,f} c_i n_i \tag{147}$$

s.t.

$$n_i \geq 0, i = 1, 2, f. \tag{148}$$

5.2 Optimal Pricing and Production Decisions

In this section, given the capacity $\boldsymbol{n}$ and realization of demand intercepts $\boldsymbol{\varepsilon}$, we characterize the optimal production vector and associated resource allocation decisions and derive the resulting revenues for two substitutable products. In other words, we solve the stage 2 problem (141) − (146), a deterministic nonlinear program. Note that for substitutable products, constraint (145) can be ignored because the prices will always be nonnegative, as will be shown later. Analysis is complicated for complementary products (i.e., $\beta < 0$) by the fact that the prices can take negative values. We partition the state space for $\boldsymbol{\varepsilon}$ into six regions corresponding to different resource

allocation choices, as in the following theorem whose proof is provided in Appendix B.

Theorem 5.2.1. *Let $\gamma_1 = \alpha_2\mu_1 - \beta\mu_2$ and $\gamma_2 = \alpha_1\mu_2 - \beta\mu_1$. Given demand realizations $\boldsymbol{\varepsilon} \geq 0$ and resource capacities $\boldsymbol{n}$ for two substitutable products (so that $\beta \geq 0$), the optimal revenue $R^*(\boldsymbol{n}, \boldsymbol{\varepsilon})$ and optimal production quantity $\boldsymbol{Q^*}$ can be uniquely determined as $R^*(\boldsymbol{n}, \boldsymbol{\varepsilon}) = \boldsymbol{Q^{*\prime}H(\varepsilon - Q^*)}$ and*

- $\boldsymbol{Q^*} = \frac{\boldsymbol{\varepsilon}}{2}$ *for* $\boldsymbol{\varepsilon} \in \Omega_1(\boldsymbol{n})$, *where*

$$\Omega_1(\boldsymbol{n}) = \Big\{\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} \geq 0; \varepsilon_1 < 2\mu_1 n_{1e}; \varepsilon_2 < 2\mu_2 n_{2e}; \frac{\varepsilon_1}{2\mu_1} + \frac{\varepsilon_2}{2\mu_2} < n_e\Big\};$$

- $Q_1^* = \frac{\varepsilon_1}{2} + \frac{\beta}{\alpha_2}(\frac{\varepsilon_2}{2} - \mu_2 n_{2e})$ *and* $Q_2^* = \mu_2 n_{2e}$ *for* $\boldsymbol{\varepsilon} \in \Omega_2(\boldsymbol{n})$, *where*

$$\Omega_2(\boldsymbol{n}) = \Big\{\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} \geq 0; \varepsilon_2 \geq 2\mu_2 n_{2e}; \varepsilon_1 + \frac{\beta}{\alpha_2}\varepsilon_2 < 2\mu_1 n_1 + \frac{\beta}{\alpha_2}2\mu_2 n_{2e}\Big\};$$

- $Q_1^* = \mu_1 n_{1e}$ *and* $Q_2^* = \frac{\varepsilon_2}{2} + \frac{\beta}{\alpha_1}(\frac{\varepsilon_1}{2} - \mu_1 n_{1e})$ *for* $\boldsymbol{\varepsilon} \in \Omega_3(\boldsymbol{n})$, *where*

$$\Omega_3(\boldsymbol{n}) = \Big\{\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} \geq 0; \varepsilon_1 \geq 2\mu_1 n_{1e}; \varepsilon_2 + \frac{\beta}{\alpha_1}\varepsilon_1 < 2\mu_2 n_2 + \frac{\beta}{\alpha_1}2\mu_1 n_{1e}\Big\};$$

- $Q_1^* = \mu_1 n_{1e}$ *and* $Q_2^* = \mu_2 n_2$ *for* $\boldsymbol{\varepsilon} \in \Omega_4(\boldsymbol{n})$, *where*

$$\Omega_4(\boldsymbol{n}) = \Big\{\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} \geq 0; \varepsilon_2 + \frac{\beta}{\alpha_1}\varepsilon_1 \geq 2\mu_2 n_2 + \frac{\beta}{\alpha_1}2\mu_1 n_{1e}; \varepsilon_1\gamma_1 - \varepsilon_2\gamma_2 \geq 2\mu_1 n_{1e}\gamma_1 - 2\mu_2 n_2\gamma_2\Big\};$$

- $Q_1^* = \mu_1 n_1$ *and* $Q_2^* = \mu_2 n_{2e}$ *for* $\boldsymbol{\varepsilon} \in \Omega_5(\boldsymbol{n})$, *where*

$$\Omega_5(\boldsymbol{n}) = \Big\{\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} \geq 0; \varepsilon_1 + \frac{\beta}{\alpha_2}\varepsilon_2 \geq 2\mu_1 n_1 + \frac{\beta}{\alpha_2}2\mu_2 n_{2e}; -\varepsilon_1\gamma_1 + \varepsilon_2\gamma_2 \geq 2\mu_2 n_{2e}\gamma_2 - 2\mu_1 n_1\gamma_1\Big\};$$

- $Q_1^* = \dfrac{\mu_1(2n_e\mu_2\gamma_2 - \varepsilon_2\gamma_2 + \varepsilon_1\gamma_1)}{2(\mu_1\gamma_1 + \mu_2\gamma_2)} > \mu_1 n_1$ *and* $Q_2^* = \dfrac{\mu_2(2n_e\mu_1\gamma_1 + \varepsilon_2\gamma_2 - \varepsilon_1\gamma_1)}{2(\mu_1\gamma_1 + \mu_2\gamma_2)} > \mu_2 n_2$
  for $\boldsymbol{\varepsilon} \in \Omega_6(\boldsymbol{n})$, *where*

$$\begin{aligned}\Omega_6(\boldsymbol{n}) \;=\; \Big\{&\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} \geq 0; \frac{\varepsilon_1}{2\mu_1} + \frac{\varepsilon_2}{2\mu_2} \geq n_e; \varepsilon_1\gamma_1 - \varepsilon_2\gamma_2 < 2\mu_1 n_{1e}\gamma_1 - 2\mu_2 n_2\gamma_2;\\ &-\varepsilon_1\gamma_1 + \varepsilon_2\gamma_2 < 2\mu_2 n_{2e}\gamma_2 - 2\mu_1 n_1\gamma_1\Big\}.\end{aligned}$$

Note that the optimum prices corresponding to each of the six cases in Theorem 5.2.1 are provided in the theorem's proof in Appendix B. We now show that whenever a firm invests in all three resource types (so that $n_1, n_2, n_f > 0$), the six regions specified in Theorem 5.2.1 are always nonempty. The proof of the following proposition can be found in Appendix B.

Proposition 5.2.1. *For any choice of model parameters and capacity choices $n_1, n_2, n_f > 0$, the regions $\Omega_1(\boldsymbol{n})$ through $\Omega_6(\boldsymbol{n})$ are not empty, and hence $P(\boldsymbol{\xi} \in \Omega_i(\boldsymbol{n})) > 0$ for $i = 1, \ldots, 6$ when ξ_1 and ξ_2 have a joint probability density function g satisfying $g(\varepsilon_1, \varepsilon_2) > 0$ for all $(\varepsilon_1, \varepsilon_2) \geq 0$.*

The regions of $\boldsymbol{\varepsilon}$ values defined in Theorem 5.2.1 corresponding to different production quantity choices are depicted in Figure 13 for $\gamma_2 > 0$. When $\gamma_2 < 0$, the slope of the parallel lines defining region $\Omega_6(\boldsymbol{n})$ will be negative. Note that if we had unlimited capacity, the optimal production decision would be $Q_i^* = \varepsilon_i/2$ for $i = 1, 2$ (see (141)). As a result of volume flexibility, the firm has the option to produce below the installed capacity. The region $\Omega_1(\boldsymbol{n})$ in Figure 13 corresponds to the unconstrained solution where we have enough capacity to produce both products optimally at $Q_i^* = \varepsilon_i/2$ for $i = 1, 2$. The region $\Omega_2(\boldsymbol{n})$ in Figure 13 corresponds to the case where all flexible capacity is assigned to product 2 and there is enough capacity to produce the first product optimally. Note that the optimum production quantity for product 1 is greater than $\varepsilon_1/2$ in this case because of the cross-price effects and the fact that product 2 cannot be produced at the level $\varepsilon_2/2$. Similarly, when $\boldsymbol{\varepsilon} \in \Omega_3(\boldsymbol{n})$, all flexible capacity is assigned to product 1, and there is enough capacity to produce product 2 optimally. On the other hand, when $\boldsymbol{\varepsilon} \in \Omega_4(\boldsymbol{n})$, then all flexible capacity is assigned to product 1, but there is not enough capacity for product 2, hence all of the dedicated capacity n_2 is used for production. Similarly, in region $\Omega_5(\boldsymbol{n})$, all flexibility is assigned to product 2, and all the dedicated capacity n_1 is used for production. Finally, when $\boldsymbol{\varepsilon} \in \Omega_6(\boldsymbol{n})$, the flexible capacity is shared between the two product groups.

Next, we consider the special cases where there is no investment in one or more of the capacity types. Then, some regions in the definition of Theorem 5.2.1 can be empty. For instance, when $n_1 = 0$, it is easy to see that $\Omega_2(\boldsymbol{n}) = \emptyset$, since we cannot produce product 1 optimally while assigning all flexible capacity to product 2. Similarly, when $n_2 = 0$ or $n_f = 0$, we have $\Omega_3(\boldsymbol{n}) = \emptyset$ and $\Omega_6(\boldsymbol{n}) = \emptyset$, respectively. Finally, when $n_1 = 0$ and $\mu_1 \neq \mu_2$, depending on the elasticity parameters, $\Omega_5(\boldsymbol{n})$ can be a singleton as stated in the following proposition.

Proposition 5.2.2. *Assume that $n_1 = 0$ and $\gamma_2 = \mu_2\alpha_1 - \mu_1\beta \leq 0$. Then $P(\boldsymbol{\xi} \in \Omega_5(\boldsymbol{n})) = 0$ when ξ_1 and ξ_2 have a joint probability density function g, implying that all flexible capacity is never assigned to product 2.*

Proof. Note that when $\gamma_2 < 0$, the two lines defining region $\Omega_5(\boldsymbol{n})$ intersect at the point $(0, 2n_{2e}\mu_2)$ and have negative slopes given by γ_1/γ_2 and $-\alpha_2/\beta$. It follows from

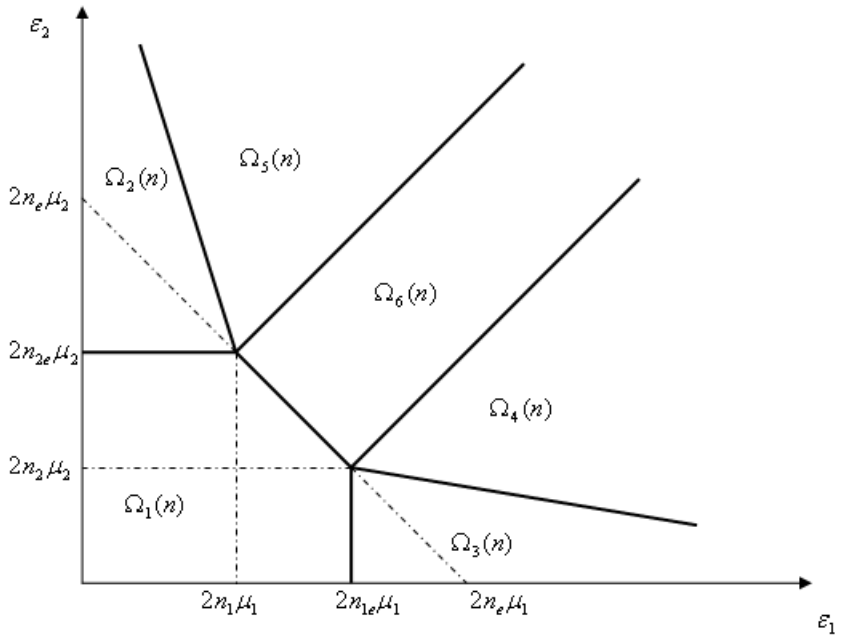


Figure 13: Characterization of optimal resource allocations with respect to demand intercepts.

(172) and $\boldsymbol{\varepsilon} \geq 0$ that $\Omega_5(\boldsymbol{n}) = \{(0, 2n_{2e}\mu_2)\}$ when $\gamma_2 < 0$. Similarly, when $\gamma_2 = 0$, the line with slope γ_1/γ_2 is vertical and overlaps the ε_2-axis, again implying that $\Omega_5(\boldsymbol{n}) = \{(0, 2n_{2e}\mu_2)\}$. In both cases, we have $P(\boldsymbol{\xi} \in \Omega_5(\boldsymbol{n})) = 0$. □

Theorem 5.2.1 generalizes earlier results of Bish and Wang [18] and Chod and Rudi [28] to the case where we have both dedicated and flexible capacities, along with product substitutability and varying price elasticities and server capabilities. By contrast, Bish and Wang [18] consider independent products ($\beta = 0$) with equal price elasticities ($\alpha_1 = \alpha_2$), where servers have the same capability at both products ($\mu_1 = \mu_2$ and $f = 1$). Chod and Rudi [28] consider substitutable products ($\beta \geq 0$), but they allow investment only in flexible servers ($n_1 = n_2 = 0$) that have the same capability at both product groups ($\mu_1 = \mu_2$) with equal price sensitivities ($\alpha_1 = \alpha_2$). Our results show that server capabilities and price elasticities have significant impact on the form of the optimal solution by affecting the region definitions, and hence where given demand intercept observations fall.

In particular, when $\beta = 0$ (the case considered in [18]), the line separating regions $\Omega_2(\boldsymbol{n})$ and $\Omega_5(\boldsymbol{n})$ is always vertical, the line separating regions $\Omega_3(\boldsymbol{n})$ and $\Omega_4(\boldsymbol{n})$ is always horizontal, and when $\alpha_1 = \alpha_2$ and $\mu_1 = \mu_2$ (the case considered in [18, 28]), the slope of the lines separating region $\Omega_6(\boldsymbol{n})$ from regions $\Omega_4(\boldsymbol{n})$ and $\Omega_5(\boldsymbol{n})$ is always

1 (because $\gamma_1/\gamma_2 = 1$). As a result, regions $\Omega_4(\boldsymbol{n}), \Omega_5(\boldsymbol{n})$, and $\Omega_6(\boldsymbol{n})$ are always unbounded, regions $\Omega_2(\boldsymbol{n})$ and $\Omega_3(\boldsymbol{n})$ are unbounded when $\beta = 0$ and $n_1, n_2 > 0$, and region $\Omega_1(\boldsymbol{n})$ is bounded. By contrast, in our case, regions $\Omega_2(\boldsymbol{n}), \Omega_3(\boldsymbol{n}), \Omega_5(\boldsymbol{n})$, and $\Omega_6(\boldsymbol{n})$ can be bounded (region $\Omega_1(\boldsymbol{n})$ is always bounded and region $\Omega_4(\boldsymbol{n})$ is always unbounded since we assume throughout the chapter that $\mu_1 \geq \mu_2$ and hence $\gamma_1 > 0$). More specifically, regions $\Omega_2(\boldsymbol{n})$ and $\Omega_3(\boldsymbol{n})$ are bounded when $\beta > 0$, and regions $\Omega_5(\boldsymbol{n})$ and $\Omega_6(\boldsymbol{n})$ are bounded when $\gamma_2 < 0$. Thus, if server capabilities are ignored, a high demand intercept value for only one product means that all flexible capacity is assigned to that product, and flexible capacity is shared between the products when both demand intercepts are simultaneously high. When we consider server capabilities and $\mu_1 > \alpha_1\mu_2/\beta$, so that μ_1 is significantly larger than μ_2 (because $\alpha_1/\beta > 1$), then $\gamma_2 < 0$, and we observe that it is possible to share the flexible capacity between the products even when ε_2 is much higher than ε_1. In fact, when $\mu_1 > \alpha_1\mu_2/\beta$, then for most values of ε_1 and ε_2, all flexible capacity is assigned to product 1. Finally, modifying the flexible capacity capability parameter f affects the parameters n_{1e}, n_{2e}, and n_e but not the slopes γ_1/γ_2, $-\beta/\alpha_1$, and $-\alpha_2/\beta$ of the lines in the definitions of the regions $\Omega_2(\boldsymbol{n}), \ldots, \Omega_6(\boldsymbol{n})$. As a result, when f decreases, the general structure of the regions remains the same, but the regions $\Omega_1(\boldsymbol{n})$ and $\Omega_6(\boldsymbol{n})$ become smaller. This is expected because reducing f means reduced total available capacity for satisfying both demands optimally (i.e., decrease in $\Omega_1(\boldsymbol{n})$) and also reduced flexible capacity to share between the products (i.e., decrease in $\Omega_6(\boldsymbol{n})$).

5.3 Optimal Capacity Decision

In this section, we analyze the capacity investment decision in the first stage under a general demand model with cross-price effects when the demand intercepts $\boldsymbol{\xi} = (\xi_1, \xi_2)'$ are discrete (or have been discretized, as in Biller, Muriel, and Zhang [13] and Lus and Muriel [67]). In other words, there are S possible demand scenarios $\boldsymbol{\varepsilon}^s = (\varepsilon_1^s, \varepsilon_2^s)'$, each occurring with probability r_s, for $s = 1, \ldots, S$. This will facilitate obtaining insights into when it is optimal to invest in flexible capacity, as well as identifying the expected values of the optimal quantities and prices.

Let $\boldsymbol{Q}^s = (Q_1^s, Q_2^s)'$ and $\boldsymbol{P}^s = (P_1^s, P_2^s)'$ be the optimum production quantities and prices in scenario $s = 1, \ldots, S$, and let $\boldsymbol{Q}$ and $\boldsymbol{\varepsilon}$ be the $S \times 2$ matrices with (s, i) entry Q_i^s and ε_i^s, respectively. Then, the objective is to jointly determine the optimal dedicated and flexible capacities, along with production levels and prices, resulting

in the following stage 1 problem

$$V^*(\boldsymbol{\varepsilon}) = \max_{\boldsymbol{Q},\boldsymbol{n}} \; V(\boldsymbol{Q},\boldsymbol{n}) = \sum_{s\in S} r_s \left[\begin{pmatrix} Q_1^s & Q_2^s \end{pmatrix} \boldsymbol{H} \begin{pmatrix} \varepsilon_1^s - Q_1^s \\ \varepsilon_2^s - Q_2^s \end{pmatrix} \right]$$
$$-c_1 n_1 - c_2 n_2 - c_f n_f \text{ s.t.} \tag{149}$$

$$Q_1^s \leq \mu_1 n_{1e}, s = 1, \ldots, S; \tag{150}$$

$$Q_2^s \leq \mu_2 n_{2e}, s = 1, \ldots, S; \tag{151}$$

$$\frac{Q_1^s}{\mu_1} + \frac{Q_2^s}{\mu_2} \leq n_e, s = 1, \ldots, S; \tag{152}$$

$$n_1, n_2, n_f \geq 0; \tag{153}$$

$$\boldsymbol{H}[\boldsymbol{\varepsilon}^s - \boldsymbol{Q}^s] \geq 0, s = 1, \ldots, S; \tag{154}$$

$$Q_1^s, Q_2^s \geq 0, s = 1, \ldots, S. \tag{155}$$

Note that this optimization problem is obtained by combining the stage 1 problem (147) − (148) with the stage 2 problem (141) − (146). Lus and Muriel [67] also considered a similar optimization problem for the specific case when $f = 1$ and $\mu_1 = \mu_2$ with general model parameters α_1, α_2, and $\beta \geq 0$ (the case with $\beta = 0$ is considered in [13]). They obtained the expected optimal quantities $E[Q_i^*] = \sum_{s=1}^S Q_i^s r_s$ and prices $E[P_i^*] = \sum_{s=1}^S P_i^s r_s$ for $i = 1, 2$ given that the firm invests either in both products (i.e., either $n_1, n_2 > 0$ or $n_f > 0$) or in only one product (i.e., either n_1 or $n_2 > 0$, $n_f = 0$). Based on the nonlinear program (149) − (155), we determine the optimal expected quantities and prices for heterogenous products and resources, show that there are only five possible scenarios in terms of optimal capacity investment, and provide a condition under which we do not invest in a given product. The proof of the following theorem is provided in Appendix B.

Theorem 5.3.1. *(a) If it is optimal for the firm to invest in only one product $i \in \{1, 2\}$ (i.e., $n_i^* > 0$ and $n_j^* = n_f^* = 0$ for $j = 3 - i$), then the expected optimal production quantities and prices are*

$$E[Q_i^*] = \frac{E[\alpha_j \xi_i + \beta \xi_j]\mu_i - c_i d}{2\alpha_j \mu_i}, \tag{156}$$

$$E[P_i^*] = \frac{E[\alpha_j \xi_i + \beta \xi_j]\mu_i + c_i d}{2d\mu_i}. \tag{157}$$

(b) Assume that $c_f/f \geq c_1, c_2$, so that flexible capacity is relatively more expensive than dedicated capacity. If it is optimal for the firm to invest in both products (i.e., either $n_1^*, n_2^* > 0$ or $n_f^* > 0$), then the optimal dedicated resource capacities n_1^*, n_2^* will always be positive. Hence, the cases where the firm invests only in the

flexible resource ($n_1^* = 0, n_2^* = 0, n_f^* > 0$)*, or in the flexible and one dedicated resource* ($n_1^* = 0$ *or* $n_2^* = 0, n_f^* > 0$) *are not possible. The expected optimal production quantities and prices when* $n_1^*, n_2^* > 0$ *are*

$$E[Q_i^*] = \frac{E[\xi_i]\mu_i\mu_j + \beta c_j\mu_i - \alpha_i c_i\mu_j}{2\mu_i\mu_j} \text{ for } i = 1, 2, i \neq j, \tag{158}$$

$$E[P_i^*] = \frac{E[\alpha_j\xi_i + \beta\xi_j]\mu_i + dc_i}{2d\mu_i} \text{ for } i = 1, 2, i \neq j. \tag{159}$$

(c) If $c_i > \mu_i E[\alpha_j\xi_i + \beta\xi_j]/d$*, then the firm should not invest in product* i *(i.e.,* $n_i^* = 0$ *and* $n_f^* = 0$*).*

Remark 5.3.1. *Through a case by case study as in the proof of Theorem 5.2.1, we can obtain the optimal quantities and prices for each demand intercept scenario, and hence the corresponding expected optimal profit. For instance, assume that the investment decision is of the form* $n_1 > 0$ *and* $n_2 = n_f = 0$*. Then there are two cases depending on whether we produce below the installed capacity or at the maximum possible capacity. Let* $\bar{\varepsilon}^s = (\alpha_2\varepsilon_1^s + \beta\varepsilon_2^s)/2\alpha_2$ *for* $s = 1, \ldots, S$*. Then we have* $Q_1^s = \bar{\varepsilon}^s$ *and* $Q_2^s = 0$ *for* $\boldsymbol{\varepsilon^s} \in \{\boldsymbol{\varepsilon^s} : \boldsymbol{\varepsilon^s} \geq 0; \bar{\varepsilon}^s < \mu_1 n_1\}$ *and* $Q_1^s = \mu_1 n_1$ *and* $Q_2^s = 0$ *for* $\boldsymbol{\varepsilon^s} \in \{\boldsymbol{\varepsilon^s} : \boldsymbol{\varepsilon^s} \geq 0; \bar{\varepsilon}^s \geq \mu_1 n_1\}$ *(this follows from equations* (184) − (185) *in Appendix B with* $Q_2^s = 0$ *and* $\lambda_3^s = 0$ *for all* s *since constraint (152) is redundant). Equation (149) now yields that the resulting profit is given by*

$$V(\boldsymbol{Q}, \boldsymbol{n}) = \frac{\alpha_2}{d}\Big(2E[\bar{\varepsilon}^s \times \min\{\bar{\varepsilon}^s, \mu_1 n_1\}] - E[(\min\{\bar{\varepsilon}^s, \mu_1 n_1\})^2]\Big) - c_1 n_1,$$

allowing for the comparison of different n_1 *values.*

In this section, we have identified the expected optimal quantities and prices for the general case with both dedicated and flexible resources and processing rates that depend on both the product and resource type, along with product substitutability and arbitrary price elasticities, under finite number of scenarios. By contrast, the previous literature only considers models where servers have similar capabilities. We have shown that if it is optimal to produce a given product, we need to invest in the corresponding dedicated capacity (as opposed to producing it exclusively using the flexible resources). This is intuitive for substitutable products because pricing can be effectively used to match demand with supply, thus reducing the need for flexible capacity. Hence, there are five possible investment strategies for substitutable products, namely $n_1^*, n_2^*, n_f^* > 0$; $n_1^*, n_2^* > 0, n_f^* = 0$; $n_1^* > 0, n_2^* = n_f^* = 0$; $n_2^* > 0, n_1^* = n_f^* = 0$; and $n_1^* = n_2^* = n_f^* = 0$, and we have identified the expected optimal

production quantity and price of each product in all cases. We observe that the price c_f and performance ratio f of the flexible capacity do not have an effect on the expected optimal production quantity and price of each product. This is consistent with the results of Lus and Muriel [67] who observe that $E[Q_1^*]$ and $E[P_i^*]$ do not depend on c_f when $\mu_1 = \mu_2$ and $f = 1$. We also observe that the expected optimal production quantity and price of each product only depend on the costs c_1, c_2 and rates μ_1, μ_2 through the effective cost ratio c_i/μ_i, $i = 1, 2$. More specifically, the expected optimal production price of a product only depends on its own effective cost ratio and increases with that parameter, while the expected optimal production quantity of a product is inversely affected by its own effective cost ratio and will increase with the effective cost ratio of the other product when it is optimal to produce both products. Finally, we have categorized when we would not produce a given product, showed that a necessary condition for investing in flexible capacity is that the costs c_1 and c_2 of both dedicated capacities be simultaneously small, and provided the bounds on c_1 and c_2 in terms of model parameters.

5.4 Numerical Analysis

In this section, we numerically study the pricing and capacity planning problem to better understand how the optimal solution depends on various model parameters. The most closely related works include Biller, Muriel, and Zhang [13] who numerically study the benefits of postponed pricing (as compared to fixed pricing) for two independent products, and Lus and Muriel [67] who conduct a numerical analysis to determine the effects of the substitutability parameter $\beta \geq 0$ in the specific case when $\mu_1 = \mu_2$, $f = 1$, $\alpha_1 = \alpha_2$, and $\sigma_1/m_1 = \sigma_2/m_2$. Others numerically study the effects of demand variance and correlation on expected capacity and profit (see, e.g., Fine and Freund [38] and Goyal, Netessine, and Randall [47]). By contrast, we study the effects of all model parameters on the optimum capacity and expected optimum production quantities and prices for a more general model. More specifically, our aim is to determine the effects of product substitutability (β), demand variability (σ_1, σ_2), correlation (ρ), server capabilities (μ_1, μ_2, f), and price sensitivity (α_1, α_2) on the optimal capacity, production decisions, and the resulting expected profits. Note that since we can revise the product prices after the planning stage to match market conditions, actual profits may differ from the expected profit.

We analyze a specific problem with a discrete set of possible values for ξ_1 and ξ_2. To determine the possible scenarios, we proceed as in Biller, Muriel, and Zhang [13]

and discretize normally distributed random variables. Note that we could discretize any continuous distribution, and the normality assumption is a convenient special case that is easily amenable to analysis with respect to means, variances, and correlation. Specifically, consider normally distributed random variables X_1 and X_2 with means m_1, m_2, standard deviations σ_1, σ_2, and correlation coefficient ρ. We discretize this normal distribution by dividing the range $(m_i \pm 2\sigma_i)$ into 10 equal intervals and using the midpoint of each interval as the value of ε_i for $i \in \{1, 2\}$ in that interval, resulting in 100 scenarios (see Biller, Muriel, and Zhang [13]). The corresponding probability for each scenario is then calculated by the probability that (X_1, X_2) falls in the corresponding range, scaled by an appropriate factor so that the total probability adds up to one. We choose the means m_1, m_2 and standard deviations σ_1, σ_2 so that all $(\varepsilon_1^s, \varepsilon_2^s)$ values obtained from the the range $(m_i \pm 2\sigma_i)$ will be positive.

In the following numerical examples, we consider two types of products. In most cases, product 1 has a larger customer base who are more price sensitive than the customers for product 2 (so that $m_1 \geq m_2$ and $\alpha_1 \geq \alpha_2$) with equal predictability ($\sigma_1 = \sigma_2$). At the same investment level, more of product 1 can be produced than product 2 (so that $\mu_1 \geq \mu_2$). Flexible servers are slower than the dedicated ones (so that $f < 1$). In all examples, given the demand scenarios and model parameters, we solve the optimization problem (149) − (155) to obtain the optimal capacities, quantities, prices, and expected profit. Note that $\beta \in (-2, 2)$ when $\alpha_1 = 3$ and $\alpha_2 = 2$ because $|\beta| < \min\{\alpha_1, \alpha_2\}$. We first conduct the analysis for independent products (i.e., $\rho = 0$), then study the effects of correlation on optimum solutions. In all cases, we let $m_1 = 500$, $m_2 = 300$, $c_1 = c_2 = 100$, and $c_f = 110$. Unless otherwise specified, we will use the following default values for our model parameters $\alpha_1 = 3$, $\alpha_2 = 2$, $\beta \in \{-1, 0, 1\}$, $\mu_1 = 15$, $\mu_2 = 10$, $f = 0.95$, $\sigma_1 = \sigma_2 = 75$, and $\rho = 0$.

The general model has many variables that interact and affect the optimal decisions differently. To be able to better understand the effects of these parameters, we consider various situations in Tables 2 through 4, where $n^* = n_1^* + n_2^* + n_f^*$ is the total capacity. In particular, we consider pairs $(\alpha_1, \alpha_2) \in \{(2, 2), (2, 3), (3, 2)\}$, and obtain the optimal solutions for $\beta \in \{-1, 0, 1\}$, $f = 0.95$, $\rho = 0$, and equal variances $\sigma_1 = \sigma_2 = 75$ when $\mu_1 = \mu_2 = 10$ in Table 2, $\mu_1 = 15$, $\mu_2 = 10$ in Table 3, and $\mu_1 = 10$, $\mu_2 = 15$ in Table 4 (relaxing the assumption that $\mu_1 \geq \mu_2$ in Table 4). This allows us to consider all possible cases with respect to the relative values of the price sensitivities α_1, α_2 under various situations.

First we look at the sensitivity of the optimal capacities and corresponding profits to the price elasticity parameters α_1 and α_2 in Tables 2-4 while other parameters are

Table 2: Sensitivity of the solution to α_1, α_2, and β with $\sigma_1 = \sigma_2 = 75$, $\mu_1 = 10$, $\mu_2 = 10$, and $\rho = 0$.

α_1	α_2	β	n_1^*	n_2^*	n_f^*	n^*	$E[Q_1^*]$	$E[Q_2^*]$	$E[P_1^*]$	$E[P_2^*]$	profit
2.00	2.00	-1.00	21.14	7.87	9.86	38.86	236.50	130.57	119.19	25.12	28696.81
2.00	3.00	-1.00	21.31	7.17	9.68	38.15	236.50	125.94	123.29	16.92	28272.52
3.00	2.00	-1.00	20.38	11.25	6.26	37.89	230.00	134.81	74.96	45.12	20375.16
2.00	2.00	0.00	22.41	12.41	5.45	40.26	240.00	140.00	130.00	80.00	39207.89
2.00	3.00	0.00	22.47	11.59	5.32	39.38	240.00	135.00	130.00	55.00	35344.62
3.00	2.00	0.00	21.59	12.47	5.32	39.38	235.00	140.00	88.33	80.00	28677.95
2.00	2.00	1.00	24.31	14.31	4.00	42.61	245.00	145.00	221.67	188.33	78574.93
2.00	3.00	1.00	24.06	13.03	4.21	41.30	245.00	140.00	185.00	115.00	58142.64
3.00	2.00	1.00	23.03	14.06	4.21	41.30	240.00	145.00	135.00	145.00	50142.64

Table 3: Sensitivity of the solution to α_1, α_2, and β with $\sigma_1 = \sigma_2 = 75$, $\mu_1 = 15$, $\mu_2 = 10$, and $\rho = 0$.

α_1	α_2	β	n_1^*	n_2^*	n_f^*	n^*	$E[Q_1^*]$	$E[Q_2^*]$	$E[P_1^*]$	$E[P_2^*]$	profit
2.00	2.00	-1.00	13.90	10.63	6.78	31.30	239.25	131.71	117.74	25.27	29541.19
2.00	3.00	-1.00	14.04	9.77	6.63	30.43	239.25	126.78	121.81	17.14	29123.73
3.00	2.00	-1.00	13.55	12.17	5.10	30.83	235.00	136.47	73.29	45.12	21175.86
2.00	2.00	0.00	14.92	13.12	4.38	32.41	243.33	140.00	128.33	80.00	40065.08
2.00	3.00	0.00	14.91	12.17	4.35	31.44	243.33	135.00	128.33	55.00	36206.98
3.00	2.00	0.00	14.51	13.16	4.29	31.96	240.00	140.00	86.67	80.00	29501.61
2.00	2.00	1.00	16.23	14.44	3.43	34.10	248.33	143.33	220.00	188.33	79475.55
2.00	3.00	1.00	15.96	13.09	3.74	32.78	248.33	138.33	183.33	115.00	59045.00
3.00	2.00	1.00	15.66	14.39	3.37	33.43	245.00	143.33	133.33	145.00	50996.45

fixed. Intuitively, more price sensitive customers means less business for the firm, as can be observed in Tables 2-4 by a decrease in the optimum profits as α_1 or α_2 increases for all β and μ_1 and μ_2 pairs. Optimum expected profit is more sensitive to product 1 elasticity than that of product 2, since mean demand for product 1 (*i.e.*, m_1) is higher than the one for product 2. For all levels of β in Tables 2-4, expected price, quantity and dedicated resource capacity for product i are decreasing with its own sensitivity parameter α_i for $i = 1, 2$. For substitutable products (i.e., $\beta \geq 0$), the change in expected prices and quantities can be explained by Theorem 5.3.1. The decrease in the dedicated resource capacity for the product for which price elasticity is increasing is a response to the fact that less of that product is required in the optimal solution. The change in the optimum dedicated resource capacity for the product whose price elasticity stays constant and change in the flexible resource capacity are less obvious and depends on whether products are substitutes or complements. For complementary products in all three tables, optimal flexible capacity is always

Table 4: Sensitivity of the solution to α_1, α_2, and β with $\sigma_1 = \sigma_2 = 75$, $\mu_1 = 10$, $\mu_2 = 15$, and $\rho = 0$.

α_1	α_2	β	n_1^*	n_2^*	n_f^*	n^*	$E[Q_1^*]$	$E[Q_2^*]$	$E[P_1^*]$	$E[P_2^*]$	profit
2.00	2.00	-1.00	22.14	3.86	8.80	34.80	238.21	133.96	119.18	23.43	29189.23
2.00	3.00	-1.00	22.25	3.51	8.69	34.44	238.21	130.92	123.26	15.27	28737.09
3.00	2.00	-1.00	21.28	7.38	5.05	33.71	231.67	138.14	74.96	43.45	20875.46
2.00	2.00	0.00	23.12	8.25	4.38	35.75	240.00	143.33	130.00	78.33	39731.75
2.00	3.00	0.00	23.16	7.84	4.29	35.29	240.00	140.00	130.00	53.33	35834.95
3.00	2.00	0.00	22.17	8.25	4.36	34.77	235.00	143.33	88.33	78.33	29206.98
2.00	2.00	1.00	24.44	9.56	3.43	37.43	243.33	148.33	221.67	186.67	79142.22
2.00	3.00	1.00	24.39	9.00	3.37	36.76	243.33	145.00	185.00	113.33	58663.11
3.00	2.00	1.00	23.09	9.29	3.74	36.12	238.33	148.33	135.00	143.33	50711.67

decreasing in all cases whenever one of the products become more sensitive to its price.

Next we analyze the effects of the server capabilities μ_1, μ_2 at the two product groups using Tables 2, 3, and 4. For all cases, we note that the results are consistent with Theorem 5.3.1 for substitutable products. That is, for $\beta \geq 0$, as μ_i increases, expected product i price is decreasing, while expected product j price is not affected. Similarly, expected product i quantity is increasing with μ_i, while expected product j quantity is decreasing for $\beta \geq 0$ and is constant for $\beta = 0$. Even though Theorem 5.3.1 only applies for substitutable products, equation (158) can be used to interpret the behavior for the complementary products such that expected product j quantity increases for $\beta < 0$ as μ_i increases. The fact that expected profit is increasing with μ_i in all cases is also expected since the effect of increasing the server capability is similar to adding some free dedicated servers. As a result, we observe that optimum dedicated resource i capacity, flexible capacity, total capacity are decreasing with increasing μ_i. This is expected since as the server becomes faster, the need for the capacities n_i^* and n_f^* is reduced and more can be produced with the same installed capacity. However, the increase in the optimum dedicated resource j capacity with an increase in μ_i is interesting to note. Even though $E[Q_2^*]$ is decreasing with μ_1 when $\beta = 1$, see Tables 2 and 3, we observe that n_2^* increases in all cases. This is in part attributed to the fact that flexible capacity n_f^* is decreasing with μ_1, hence to cope with that n_2^* increases.

We also conducted a similar analysis for the case when $\mu_1 = \mu_2 = 10$ and the variances for the two products are not equal. In particular, we let $\sigma_1 = 150$, $\sigma_2 = 75$ in Table 5 and $\sigma_1 = 75$, $\sigma_2 = 150$ in Table 6. We choose the same service rate to isolate and study the effect of variance better. By comparing Tables 5 and 6

with Table 2, we observe that for substitutable products, increasing the variability of the demand intercept for product i results in increased investment for the resource dedicated to that product and the flexible resource, but decreased investment in the resource dedicated to the other product. The increase in investment for dedicated resource n_i is intuitive because the firm does not want to miss the opportunity to meet higher demand levels. Since the difference between the demand intercept realizations are higher, flexible capacity becomes more attractive. Similarly, the decrease in the investment for dedicated resource n_j is intuitive because given the increase in capacity of dedicated resource n_i, more of the flexible server time now can be spared for product j. As expected, the optimal prices and quantities for substitutable products are not affected, as explained in Theorem 5.3.1. Finally, we obtain a new insight from Tables 5 and 6 that for complementary products (i.e., $\beta < 0$), it is possible to invest in flexible capacity even if the optimal dedicated resource 2 capacity is zero. Hence Theorem 5.3.1 (a) does not necessarily hold when $\beta < 0$. This is expected because pricing is a less effective tool for shifting demand to match the available capacity for complementary products than for substitutable products (increasing the price of one product inversely affects the demand of the other), increasing the need for flexible capacity.

Table 5: Sensitivity of the solution to α_1, α_2, and β with $\sigma_1 = 150$, $\sigma_2 = 75$, $\mu_1 = 10$, $\mu_2 = 10$, and $\rho = 0$.

α_1	α_2	β	n_1^*	n_2^*	n_f^*	n^*	$E[Q_1^*]$	$E[Q_2^*]$	$E[P_1^*]$	$E[P_2^*]$	profit
2.00	2.00	-1.00	25.53	0.00	19.45	44.97	243.69	134.62	115.75	24.82	30177.35
2.00	3.00	-1.00	26.11	0.00	18.83	44.94	244.47	133.22	119.96	15.61	29564.69
3.00	2.00	-1.00	21.53	9.52	8.81	39.86	230.00	134.37	74.88	45.38	21340.74
2.00	2.00	0.00	24.46	10.81	7.79	43.06	240.00	140.00	130.00	80.00	40446.80
2.00	3.00	0.00	24.13	9.57	8.21	41.91	240.00	135.00	130.00	55.00	36598.34
3.00	2.00	0.00	23.24	11.09	7.43	41.76	235.00	140.00	88.33	80.00	29400.86
2.00	2.00	1.00	27.31	13.18	5.63	46.12	245.00	145.00	221.67	188.33	80301.66
2.00	3.00	1.00	26.43	11.30	6.73	44.46	245.00	140.00	185.00	115.00	59675.83
3.00	2.00	1.00	25.59	13.05	5.66	44.29	240.00	145.00	135.00	145.00	51039.55

Next, we investigate the effect of the substitutability/complementary parameter $\beta \in \{-1.8, -1.6, \ldots, +1.8\}$ on the optimum solution for independent products ($\rho = 0$) with price elasticities $\alpha_1 = 3$ and $\alpha_2 = 2$ in Table 7. Since the parameter β has the biggest impact on the optimal profits, we choose a wide range of possible β values for independent products to remove any affect from correlation. Note that as β increases, products gradually change from most complementary to most substitutable. The effect of β on the optimal flexible capacity and profits are consistent with the

Table 6: Sensitivity of the solution to α_1, α_2, and β with $\sigma_1 = 75$, $\sigma_2 = 150$, $\mu_1 = 10$, $\mu_2 = 10$, and $\rho = 0$.

α_1	α_2	β	n_1^*	n_2^*	n_f^*	n^*	$E[Q_1^*]$	$E[Q_2^*]$	$E[P_1^*]$	$E[P_2^*]$	profit
2.00	2.00	-1.00	32.61	0.00	26.20	58.81	259.95	136.71	105.60	28.85	27529.54
2.00	3.00	-1.00	33.44	0.00	25.33	58.77	259.95	135.92	111.22	17.62	26458.35
3.00	2.00	-1.00	29.65	0.00	26.16	55.81	254.80	144.61	67.00	44.20	20225.21
2.00	2.00	0.00	20.81	14.46	7.79	43.06	240.00	140.00	130.00	80.00	40446.80
2.00	3.00	0.00	21.09	13.24	7.43	41.76	240.00	135.00	130.00	55.00	36067.53
3.00	2.00	0.00	19.57	14.13	8.21	41.91	235.00	140.00	88.33	80.00	29931.68
2.00	2.00	1.00	23.18	17.31	5.63	46.12	245.00	145.00	221.67	188.33	80301.66
2.00	3.00	1.00	23.05	15.59	5.66	44.29	245.00	140.00	185.00	115.00	59039.55
3.00	2.00	1.00	21.30	16.43	6.73	44.46	240.00	145.00	135.00	145.00	51675.83

results of Birge, Drogosz, and Duenyas [14], Biller, Muriel, and Zhang [13], Bish and Suwandechochai [17], and Lus and Muriel [67] in that the optimal flexible capacity is decreasing in β while the optimal profit is increasing. Another striking change is observed in the choice of dedicated vs flexible capacities. Even though we observe an increase in the total capacity of 17.7%, the flexible capacity decreases from 39.8% of the total to 0.06% as β increases. Thus the increase in the total capacity investment is attributed to an increase in both dedicated capacities. This is expected because as the products become more substitutable, pricing can be used effectively to shift the demand from one product to the other to fit the available fixed capacity, hence reducing the need for the flexible capacity. However, the most striking change is observed in the prices P_1, P_2 and the expected profit. We observe a total increase of 269% for the price of product 1, 1462% for the price of product 2, and 528% for the expected profit. This is expected and can be explained mathematically by the form of demand functions (132) and (133). In particular, the demand D_i for product i is affected by the price of the other product P_j by an amount βP_j. If β is negative (i.e., complementary), increasing the price P_j, inversely affects the demand for product i, D_i. Hence, we are constrained to choose low prices for both products, resulting in low profit. On the other hand, when β is positive (i.e., substitutable), increasing the price P_j for product j has a positive effect on the demand for product i, D_i, resulting in higher prices and expected profit.

Next, we would like to study the effects of improving the capability of the flexible servers from 60% of the dedicated capacities to 100% of the dedicated ones (i.e., $f \in \{0.6, 0.8, 1\}$) in Table 8. Other parameters are at their default values except the variances. We choose the maximum possible variances for both products $\sigma_1 = 250$ and $\sigma_2 = 150$ so that flexible capacity becomes an attractive option and the parameter f

Table 7: Sensitivity of the solution to β with $\sigma_1 = \sigma_2 = 75$, $\alpha_1 = 3$, $\alpha_2 = 2$, $\mu_1 = 15$, $\mu_2 = 10$, and $\rho = 0$.

β	n_1^*	n_2^*	n_f^*	n^*	$E[Q_1^*]$	$E[Q_2^*]$	$E[P_1^*]$	$E[P_2^*]$	profit
-1.80	12.66	5.35	11.94	29.94	232.71	120.13	76.38	21.19	18594.18
-1.60	12.99	8.18	9.18	30.35	233.59	127.62	74.71	26.42	18978.73
-1.40	13.22	11.44	5.83	30.49	233.57	132.10	73.71	32.35	19541.37
-1.20	13.37	12.01	5.24	30.63	234.00	134.86	73.21	38.65	20258.12
-1.00	13.55	12.17	5.10	30.83	235.00	136.47	73.29	45.12	21175.86
-0.80	13.75	12.34	4.95	31.03	236.00	137.33	74.23	51.64	22310.17
-0.60	13.94	12.51	4.79	31.24	237.00	138.00	76.03	58.19	23677.30
-0.40	14.12	12.71	4.64	31.47	238.00	138.67	78.68	64.93	25301.66
-0.20	14.30	12.93	4.48	31.71	239.00	139.33	82.19	72.11	27223.24
0.00	14.51	13.16	4.29	31.96	240.00	140.00	86.67	80.00	29501.61
0.20	14.72	13.39	4.11	32.22	241.00	140.67	92.26	88.89	32221.87
0.40	14.93	13.61	3.95	32.49	242.00	141.33	99.22	99.18	35505.47
0.60	15.16	13.85	3.77	32.78	243.00	142.00	107.94	111.38	39528.72
0.80	15.40	14.11	3.58	33.09	244.00	142.67	119.01	126.27	44555.44
1.00	15.66	14.39	3.37	33.43	245.00	143.33	133.33	145.00	50996.45
1.20	15.95	14.72	3.13	33.80	246.00	144.00	152.46	169.47	59526.97
1.40	16.26	15.07	2.88	34.21	247.00	144.67	179.08	203.02	71338.15
1.60	16.61	15.47	2.60	34.68	248.00	145.33	218.45	252.09	88743.96
1.80	17.09	15.94	2.22	35.24	249.00	146.00	282.32	331.09	116910.65

has a stronger effect on the optimal solution. As expected, for substitutable products (i.e., $\beta \geq 0$), as the flexible server capability improves, investment in flexible servers increases while investment in the dedicated capacities decreases. The increase in resulting profits is also expected since we can produce more at the same investment level if needed, as the server capability improves. The change in expected quantities and prices for both types of products can be explained by Theorem 5.3.1.

Next, we look at the sensitivity of the optimal capacities and corresponding profits with respect to demand variability. For this case we assume $\sigma_1 = \sigma_2 = \sigma$ and all the other parameters are at their default values. To see the effect of demand variability on capacity choices, we generate demand scenarios for independent products with $\sigma_1 = \sigma_2 = \sigma \in \{25, 75, 125, 150\}$ and obtain the optimal solutions for $\beta \in \{-1, 0, 1\}$ in Table 9. Variances are chosen so that nonnegativity of the demand intercept scenarios is preserved. For all levels of β values, we observe, as expected, that expected profit, total flexible capacity, and total capacity are increasing with σ. This is consistent with the existing literature (see, e.g., Chod and Rudi [28], Fine and Freund [38], Goyal and Netessine [44, 45], and Lus and Muriel [67]) and results from the fact that extra

Table 8: Sensitivity of the solution to f and β with $\sigma_1 = 250$, $\sigma_2 = 150$, $\alpha_1 = 3$, $\alpha_2 = 2$, $\mu_1 = 15$, $\mu_2 = 10$, and $\rho = 0$.

f	β	n_1^*	n_2^*	n_f^*	n^*	$E[Q_1^*]$	$E[Q_2^*]$	$E[P_1^*]$	$E[P_2^*]$	profit
0.60	-1.00	10.33	0.00	78.33	88.67	256.05	146.58	66.90	43.26	20340.51
0.60	0.00	23.29	18.94	0.96	43.18	240.00	140.00	86.67	80.00	33422.92
0.60	1.00	24.31	20.07	0.00	44.38	245.00	143.33	133.33	145.00	56023.28
0.80	-1.00	10.33	0.00	58.75	69.08	256.05	146.58	66.90	43.26	22494.68
0.80	0.00	17.80	13.90	9.43	41.12	240.00	140.00	86.67	80.00	33613.23
0.80	1.00	20.35	16.24	6.84	43.43	245.00	143.33	133.33	145.00	56118.41
1.00	-1.00	10.33	0.00	47.00	57.33	256.05	146.58	66.90	43.26	23787.18
1.00	0.00	14.06	10.35	13.63	38.04	240.00	140.00	86.67	80.00	33893.16
1.00	1.00	16.42	12.78	11.68	40.88	245.00	143.33	133.33	145.00	56343.97

revenue when demand and prices are high dominates the loss in revenue when demand and prices are low. To deal with high levels of demand, the total capacity and flexible capacity are increasing with demand variance for any β. Effects on other parameters depend on whether the products are substitutes or complementary. When $\beta \geq 0$, expected prices and quantities do not depend on the demand variance, as shown in Theorem 5.3.1. As in Tables 5 and 6, we observe that the result of Theorem 5.3.1 (a) does not hold for $\beta < 0$, since we have $n_2^* = 0$ while $n_f^* > 0$ for $\sigma = 125$ and $\beta = -1$.

Table 9: Sensitivity of the solution to σ and β with $\sigma_1 = \sigma_2 = \sigma \in \{25, 75, 125, 150\}$, $\alpha_1 = 3$, $\alpha_2 = 2$, $\mu_1 = 15$, $\mu_2 = 10$, and $\rho = 0$.

σ	β	n_1^*	n_2^*	n_f^*	n^*	$E[Q_1^*]$	$E[Q_2^*]$	$E[P_1^*]$	$E[P_2^*]$	profit
25	-1.00	14.99	13.03	1.44	29.45	235.00	136.67	73.33	45.00	20504.12
75	-1.00	13.55	12.17	5.10	30.83	235.00	136.47	73.29	45.12	21175.86
125	-1.00	12.82	0.00	21.59	34.41	238.94	135.15	71.46	46.70	22433.49
150	-1.00	20.69	0.00	25.95	46.63	255.08	143.78	66.72	44.75	22347.28
25	0.00	15.50	13.56	1.10	30.16	240.00	140.00	86.67	80.00	29025.06
75	0.00	14.51	13.16	4.29	31.96	240.00	140.00	86.67	80.00	29501.61
125	0.00	13.81	13.40	7.51	34.73	240.00	140.00	86.67	80.00	30704.76
150	0.00	13.56	13.52	9.21	36.30	240.00	140.00	86.67	80.00	31592.54
25	1.00	16.09	14.16	0.73	30.98	245.00	143.33	133.33	145.00	50408.68
75	1.00	15.66	14.39	3.37	33.43	245.00	143.33	133.33	145.00	50996.45
125	1.00	15.52	15.20	6.07	36.79	245.00	143.33	133.33	145.00	52478.69
150	1.00	15.50	15.67	7.47	38.64	245.00	143.33	133.33	145.00	53568.30

Finally, we examine the sensitivity of the optimal capacities and corresponding profits with respect to demand correlation. In particular, we generate demand scenarios for $\rho \in \{-0.5, +0.5\}$ and obtain the optimal solutions for $(\alpha_1, \alpha_2) \in \{(2,2), (2,3), (3,2)\}$

and $\beta \in \{-1, 0, 1\}$ in Tables 10 and 11, respectively. Even though correlation is comparable to β in terms of its effects on the optimal flexible capacity, it has only a slight effect on the expected profits. We observe that as the product correlation increases, the optimum flexible capacity decreases while total capacity investment increases. Even though the need for capacity increases to take advantage of a potentially large market as the products become more correlated, the value of the flexible resource is reduced because relative values of demand for both products become more predictable as the correlation ρ increases. We do not observe a substantial effect of demand correlation on the optimal profit. For $\beta \geq 0$, we note that the expected prices for products 1 and 2, and expected quantities remain the same as ρ increases. This is expected since, as we obtained in Theorem 5.3.1, expected prices and quantities do not depend on the demand correlation.

Table 10: Sensitivity of the solution to α_1, α_2, and β with $\mu_1 = 15$, $\mu_2 = 10$, $\sigma_1 = \sigma_2 = 75$, and $\rho = -0.5$.

α_1	α_2	β	n_1^*	n_2^*	n_f^*	n^*	$E[Q_1^*]$	$E[Q_2^*]$	$E[P_1^*]$	$E[P_2^*]$	profit
2.00	2.00	-1.00	12.89	9.90	7.84	30.63	239.12	129.85	117.20	26.48	29817.82
2.00	3.00	-1.00	13.16	9.22	7.55	29.93	239.12	125.21	121.57	17.74	29311.01
3.00	2.00	-1.00	12.66	11.27	6.43	30.36	235.00	136.26	73.25	45.24	21396.22
2.00	2.00	0.00	13.87	12.10	5.74	31.71	243.33	140.00	128.33	80.00	40127.38
2.00	3.00	0.00	13.96	11.27	5.63	30.85	243.33	135.00	128.33	55.00	36264.33
3.00	2.00	0.00	13.52	12.18	5.62	31.32	240.00	140.00	86.67	80.00	29563.53
2.00	2.00	1.00	15.03	13.30	4.80	33.12	248.33	143.33	220.00	188.33	79251.71
2.00	3.00	1.00	14.89	12.15	4.94	31.98	248.33	138.33	183.33	115.00	58935.92
3.00	2.00	1.00	14.49	13.27	4.79	32.55	245.00	143.33	133.33	145.00	50892.46

Table 11: Sensitivity of the solution to α_1, α_2, and β with $\mu_1 = 15$, $\mu_2 = 10$, $\sigma_1 = \sigma_2 = 75$, and $\rho = 0.5$.

α_1	α_2	β	n_1^*	n_2^*	n_f^*	n^*	$E[Q_1^*]$	$E[Q_2^*]$	$E[P_1^*]$	$E[P_2^*]$	profit
2.00	2.00	-1.00	15.02	11.22	5.59	31.83	239.36	133.95	118.41	23.82	29217.85
2.00	3.00	-1.00	15.10	10.30	5.51	30.91	239.36	128.96	122.18	16.29	28904.75
3.00	2.00	-1.00	14.59	13.15	3.44	31.18	235.00	136.63	73.33	45.02	20926.42
2.00	2.00	0.00	16.01	14.10	2.82	32.93	243.33	140.00	128.33	80.00	39980.30
2.00	3.00	0.00	15.96	13.04	2.86	31.86	243.33	135.00	128.33	55.00	36132.18
3.00	2.00	0.00	15.58	14.20	2.65	32.42	240.00	140.00	86.67	80.00	29425.31
2.00	2.00	1.00	17.49	15.43	1.89	34.80	248.33	143.33	220.00	188.33	79664.61
2.00	3.00	1.00	17.08	13.92	2.38	33.38	248.33	138.33	183.33	115.00	59131.11
3.00	2.00	1.00	16.83	15.45	1.78	34.05	245.00	143.33	133.33	145.00	51082.31

Note that in the examples for substitutable products, we always observe $n_1^* > 0$ and $n_2^* > 0$, i.e., in all cases there is investment in both dedicated capacities. This is

due to the way we construct the demand scenarios since the minimum value for both demand intercepts is greater than 0 for both products with probability 1.

In this section, we observe that taking the server capabilities into account has an impact on the form of optimal capacity investment decisions as well as the optimal expected prices and output quantities. Similarly, the fact that flexible servers might be slower than the dedicated ones also reduces the value of the flexible servers, which in turn affects the form of optimal solutions.

5.5 Conclusion

In this chapter, we have studied a two stage stochastic capacity sizing and pricing problem for a two product firm under a linear demand model with substitutability and with an emphasis on the differences in the service capacities of flexible and dedicated resources. We formulated nonlinear programs to determine the optimal production, pricing, and resource allocation decisions given the available capacity, and derived the resulting revenues as functions of demand intercepts in six regions. We showed that taking the server capabilities into account significantly affects the form of these regions and hence the resulting optimal allocations. For instance, we identified cases where flexible capacity is never assigned to one product for any demand realization; such cases are not possible when the service rates are assumed to be equal.

For the specific case with a discrete demand intercept distribution, we showed that there are only five possible capacity investment scenarios that can be optimal and we identified the expected optimal quantities and prices corresponding to each of these scenarios. We concluded that the expected optimal prices and quantities are determined by the effective cost ratio of the dedicated servers, defined as the ratio of the server's cost to its service rate, and that the flexible server's cost and performance ratio have no effect. We also showed that investment in the resource dedicated to a given product is a prerequisite for the production of that product to be optimal, and provided an upper bound on the dedicated server's cost that should be satisfied if we are to invest in the corresponding product.

Finally, through numerical examples, we studied the effects of various model parameters on the optimal capacity, pricing, and quantity decisions, as well as the expected optimal profits. We showed that the results for substitutable products do not necessarily hold for complementary products and that server capabilities have a significant effect on the form of the optimal solution by affecting the capacity decision, as well as the resulting expected quantities, prices, and profits.

CHAPTER VI

SUMMARY AND DIRECTIONS FOR FUTURE RESEARCH

In this chapter we summarize the major contributions of this dissertation and suggest possibilities for future research. For more detailed information, the reader is referred to Chapters 3, 4, and 5.

In Chapter 3, we studied the effects of server flexibility for a multi-class queueing network that can be unstable. We allowed multiple arrival streams, as well as servers who cooperate or work in parallel when multiple servers are assigned to a class. Moving a server is also assumed to incur a random switching time that can depend on the origin and destination. We showed that the classes can be uniquely classified into stable and unstable sets and also developed server allocation policies that can achieve throughput arbitrarily close to the maximum throughput achievable given sufficient offered demand. Similarly, we provided the minimum offered demand required to achieve a feasible target throughput. Our numerical results suggest that system throughput can be significantly improved by allowing instability.

Another performance measure of interest is the total number of items processed during each server visit to a given class (i.e., the lot sizes). In general, low switching rates are effective with respect to throughput, but they can result in the production of large lots, which in turn implies longer lead times and higher inventories. Hence, in future work it would be interesting to design policies that simultaneously consider throughput and lot sizes.

In Chapter 4, we studied the effects of inspection location decisions on product quality and quantity for a general model with multiple defect types, defect dependent inspection errors, fractional inspection, probabilistic repair stations, scraps at inspection and repair stations, and stochastic costs. Our model is more general than any model considered in the inspection location literature, and also incorporates the system capacity into the inspection location determination. More specifically, we analyzed the defect propagation and flow of parts through each system stage sequentially, and also showed how to obtain the long-run profit rate for the system. Considering the general case where any of the stations can be the bottleneck, we developed an admission control policy that results in cost reduction for given inspection locations

and levels, and also introduced methods for determining the optimal inspection locations and levels. Finally, we provided numerical results that show how the inspection location decisions are made under different parameter values for a system with two production stations. We demonstrated that taking bottleneck considerations into account when determining the best inspection locations can lead to different inspection decisions than do previous models (that do not take the capacity of the system into account).

In some systems, defect causes in the production line can be traced. Then another method for quality improvement is to stop the production line until a source of defects is removed from the system, instead of scrapping or repairing the defective units, also known as "continuous improvement." It would be a nice future research topic to study the effects of inspection on system capacity under our inspection model framework when continuous improvement is employed.

Finally, in Chapter 5, we analyzed the capacity and pricing decisions made by a monopolistic firm producing two heterogenous products under demand uncertainty. The objective is to maximize the firm's profit. Our model incorporates dedicated and flexible resources, product substitutability, and processing rates that depend on the product and resource type. We provided the optimum prices and production quantities as functions of resource capacities and demand intercepts, and showed that incorporating server-dependent processing rates results in a significant shift towards assigning flexible servers to the product that can be produced faster. When the demand uncertainty has only a finite set of possible values, we showed that there are five possible capacity investment scenarios that can be optimal and identified the expected optimal quantities and prices for each of these scenarios. We also showed that investment in flexible capacity is only desirable when it is optimal to invest in dedicated capacities for both products. We concluded with numerical examples that provide insights into how the optimal capacities and expected production quantities, prices, and profit depend on various model parameters.

In our problem, we assumed that it is possible for the firm to install fractional capacity. However, in real-life applications, installed capacity might be constrained to be integer (as in an integer number of machines). It would be an interesting future research topic to optimize discrete capacity. Also, the effects of demand variance and correlation on the optimum capacities, prices, and profit could be studied rigorously for the special case where the firm only has flexible capacity.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Before proving Theorem 4.4.1, we need a preliminary result about the relation between the incoming fraction of defective units at different stations and the system input λ.

Proposition A.0.1. *Under Assumptions 4.4.1 and 4.4.2, let $U_R(\lambda)$ be the set of stages i with an unstable repair station under input rate λ. Suppose $\lambda_1 \leq \lambda_2 \leq \lambda$. Then we have $U_R(\lambda_1) \subseteq U_R(\lambda_2)$. Moreover, $\pi^O_{i,j}$, $\pi^I_{i,j}$, and $\pi^R_{i,j}$ are nondecreasing in λ_2 if $j \in \cup_{i' \in U_R(\lambda), i > i'} D_{i'}$ and are constant in λ_2 otherwise.*

Proof. First observe that defect fraction propagation only depends on the system rates when outputs of inspection and repair stations join, as shown in Section 4.2.4. Hence instability of inspection and operation stations has no effect on the defect structure (due to the use of the FCFS discipline). From the flow equations in Section 4.2, it is easy to see that flow rates to each stage λ_i are nondecreasing in the input rate λ to the system. Moreover, by equations (71) and (77), the outflow rate λ^I_i from I_i is also nondecreasing in λ. From equations (78), (79), (88), (89), and (91), we see that if R_i is stable, then as λ^I_i increases, so do λ^{IO}_i and λ^{RO}_i proportionally, and $\pi^O_{i+1,j}$ is not affected. However, if R_i becomes unstable before I_i, then as λ^I_i increases, $\pi^O_{i+1,j}$ will eventually get closer to $\pi^{IO}_{i,j}$.

We analyze the system in a step by step manner starting with the first bottleneck repair station. Let l_λ be the first stage with a repair station that is unstable under the input rate λ to the production line. Next we show that $\pi^{RO}_{l_\lambda,j} \leq \pi^{IO}_{l_\lambda,j}$ for $j \in D$. For defect types $j \in D \setminus D_{l_\lambda}$, we already have $\pi^{RO}_{l_\lambda,j} = \pi^{IO}_{l_\lambda,j} = \pi^I_{l_\lambda,j}$ by equations (83) and (90). The only way a unit routed from R_{l_λ} to $O_{l_\lambda+1}$ will have defect $j \in D_{l_\lambda}$ is when this defect is not detected by the inspection station I_{l_λ}. Otherwise, it would be either repaired by R_{l_λ} or scrapped by R_{l_λ} or I_{l_λ}. Let $u^R_{l_\lambda,j}$ and $u^O_{l_\lambda,j}$ be the fractions of units routed to R_{l_λ} and $O_{l_\lambda+1}$ from I_{l_λ} that have an undetected defect $j \in D_{l_\lambda}$, respectively. Then we have

$$u^R_{l_\lambda,j} = \frac{\pi^I_{l_\lambda,j}\beta_{l_\lambda,j}r^I_{l_\lambda,j}}{r^I_{l_\lambda}} = \begin{cases} \dfrac{\pi^I_{l_\lambda,j}\beta_{l_\lambda,j}[1-\prod_{k\in D^R_{l_\lambda}\setminus\{j\}}(1-d_{l_\lambda,k})]}{1-\prod_{k\in D^R_{l_\lambda}}(1-d_{l_\lambda,k})}, & j\in D^R_{l_\lambda}, \\ \dfrac{\pi^I_{l_\lambda,j}\beta_{l_\lambda,j}\prod_{k\in D^S_{l_\lambda}\setminus\{j\}}(1-d_{l_\lambda,k})}{\prod_{k\in D^S_{l_\lambda}}(1-d_{l_\lambda,k})}, & j\in D^S_{l_\lambda}, \end{cases}$$

and

$$\begin{aligned} u^O_{l_\lambda,j} &= \frac{\pi^I_{l_\lambda,j}[1-f_{l_\lambda}+f_{l_\lambda}\beta_{l_\lambda,j}\prod_{k\in D_{l_\lambda}\setminus\{j\}}(1-d_{l_\lambda,k})]}{o^I_l} \\ &= \frac{\pi^I_{l_\lambda,j}[1-f_{l_\lambda}+f_{l_\lambda}\beta_{l_\lambda,j}\prod_{k\in D_{l_\lambda}\setminus\{j\}}(1-d_{l_\lambda,k})]}{\prod_{k\in D_{l_\lambda}}(1-d_{l_\lambda,k})}, j\in D_{l_\lambda}. \end{aligned}$$

To see the relations between the above defect probabilities, let $M^R_{l_\lambda}=\prod_{k\in D^R_{l_\lambda}}(1-d_{l_\lambda,k})$, $M^S_{l_\lambda}=\prod_{k\in D^S_{l_\lambda}}(1-d_{l_\lambda,k})$, $M^R_{l_\lambda,j}=\prod_{k\in D^R_{l_\lambda}\setminus\{j\}}(1-d_{l_\lambda,k})$, and finally $M^S_{l_\lambda,j}=\prod_{k\in D^S_{l_\lambda}\setminus\{j\}}(1-d_{l_\lambda,k})$. Then we get

$$u^R_{l_\lambda,j} = \begin{cases} \dfrac{\pi^I_{l_\lambda,j}\beta_{l_\lambda,j}[1-M^R_{l_\lambda,j}]}{1-M^R_{l_\lambda}}, & j\in D^R_{l_\lambda}, \\ \dfrac{\pi^I_{l_\lambda,j}\beta_{l_\lambda,j}M^S_{l_\lambda,j}}{M^S_{l_\lambda}}, & j\in D^S_{l_\lambda}, \end{cases} \tag{160}$$

and

$$u^O_{l_\lambda,j} = \begin{cases} \dfrac{\pi^I_{l_\lambda,j}[1-f_{l_\lambda}+f_{l_\lambda}\beta_{l_\lambda,j}M^R_{l_\lambda,j}M^S_{l_\lambda}]}{M^R_{l_\lambda}M^S_{l_\lambda}}, & j\in D^R_{l_\lambda}, \\ \dfrac{\pi^I_{l_\lambda,j}[1-f_{l_\lambda}+f_{l_\lambda}\beta_{l_\lambda,j}M^R_{l_\lambda}M^S_{l_\lambda,j}]}{M^R_{l_\lambda}M^S_{l_\lambda}}, & j\in D^S_{l_\lambda}. \end{cases} \tag{161}$$

Since $M^R_{l_\lambda}\le M^R_{l_\lambda,j}$, $M^R_{l_\lambda,j}M^S_{l_\lambda}\le 1$, and $M^R_{l_\lambda}M^S_{l_\lambda,j}\le 1$, we have

$$\frac{1-M^R_{l_\lambda,j}}{1-M^R_{l_\lambda}} \le \frac{M^R_{l_\lambda,j}}{M^R_{l_\lambda}} \text{ and } u^O_{l_\lambda,j} \ge \begin{cases} \dfrac{\pi^I_{l_\lambda,j}\beta_{l_\lambda,j}M^R_{l_\lambda,j}}{M^R_{l_\lambda}}, & j\in D^R_{l_\lambda}, \\ \dfrac{\pi^I_{l_\lambda,j}\beta_{l_\lambda,j}M^S_{l_\lambda,j}}{M^S_{l_\lambda}}, & j\in D^S_{l_\lambda}. \end{cases} \tag{162}$$

By equations (160) and (162), we now obtain $u^R_{l_\lambda,j}\le u^O_{l_\lambda,j}$ for all $j\in D$. This means that the units routed to repair stations have fewer undetected defects than units routed to operation stations. Since repair stations do not introduce any defects and units routing from a repair station to the next operation station do not have any detected defects, this implies $\pi^{RO}_{l_\lambda,j}\le\pi^{IO}_{l_\lambda,j}$ for $j\in D$. Hence, as λ_2 increases and $R_{l_{\lambda_2}}$

becomes unstable, the defect fractions to the next operation station $\pi^O_{l_\lambda+1,j}$ for $j \in D_{l_\lambda}$ is nondecreasing, while $\pi^O_{l_\lambda+1,j}$ for $j \in D \setminus D_{l_\lambda}$ remains unaffected.

Note that even though the input to the system is increased, the routing probabilities out of all stations in all stages are unaffected. This is due to the fact that routing probabilities in a given stage i depend only on the fraction of units having the set of defects D_i, and by Assumption 4.4.2, this fraction cannot be modified at any stage other than stage i. Even if this fraction is modified at stage i by an unstable repair station R_i, this occurs only after the stage i is completed for parts that already left stage i. Then it is easy to see that $U_R(\lambda_1) \subseteq U_R(\lambda_2)$ for $\lambda_1 \leq \lambda_2$, since flow through all stations increases proportionally as the input flow is increased.

After stage l_λ, there can be either no more bottleneck repair stations or one or more bottleneck repair stations. In the first case, the defect fractions for $j \in D \setminus D_{l_\lambda}$ stay constant throughout the production line. In the second case, let l'_λ be the second stage with a bottleneck repair station under λ. Similar to the analysis in the l_λth stage, the fraction of units having defect $j \in D_{l'_\lambda}$ is nondecreasing throughout stages $i > l'_\lambda$. Continuing sequentially, it is easy to see that $\pi^O_{i,j}$, $\pi^I_{i,j}$, and $\pi^R_{i,j}$ are nondecreasing in λ_2 if $j \in \cup_{i' \in U_R(\lambda), i > i'} D_{i'}$ and are constant otherwise. □

We are now ready to prove Theorem 4.4.1.

Proof of Theorem 4.4.1. We show, by contradiction, that under the optimal policy, no operation or inspection station can have $\lambda_i > \lambda^O_i$ or $\lambda^O_i > \lambda^I_i$, respectively, because we can always improve the total profit by stabilizing the operation and inspection stations in the production line. For this, assume that there exists a stage i such that $\lambda_i > \lambda^O_i$ and/or $\lambda^O_i > \lambda^I_i$, and let stage b be the first such stage. We start by considering the case when $\lambda_b > \lambda^O_b$, so that after O_b, the arrival rate to inspection station I_b is $\lambda^O_b = \mu^O_b$. Later, we discuss the case when $\lambda_b \leq \lambda^O_b$ and $\lambda^O_b > \lambda^I_b$, so that the first bottleneck is an inspection station.

Note that each unit generates a revenue of R only if it reaches the end of the line. Otherwise, it incurs various nonnegative costs throughout the serial line. Let λ be the current input rate to the system and $\lambda_1 < \lambda$ be the smallest input rate to operation station O_1 such that $\lambda_b = \lambda^O_b = \mu^O_b$. Note that the inflow rate to operation station O_b is a continuous and nondecreasing function of λ and is null when $\lambda = 0$ (see Section 4.2), implying that λ_1 exists. We compare the system with input rate λ_1 to the system with input rate λ. For this, we study the effects of increasing the arrival rate $\lambda_2 \in [\lambda_1, \lambda]$ on the first part of the production line (up to the bottleneck station O_b) and on the second part (after the bottleneck station O_b).

Note that increasing the input flow not only affects the flow of parts at various stations in the line, but also the quality characteristics of the products. Let l_λ be the first stage with a repair station that is unstable under the input rate λ. We first consider the case with $l_\lambda \geq b$, so that there are no bottleneck repair stations before stage b, and then the case with $l_\lambda < b$, so that there is one or more bottleneck repair station before the bth stage.

If $l_\lambda \geq b$, then, by Proposition A.0.1 and Assumption 4.4.2, the fraction of units having defects in D_i for $i < b$, is constant in the input rate λ. Since routing probabilities out of all stations in a given stage i depend only on the set of defects D_i and are not affected by the set of defects $D \setminus D_i$, the routing probabilities at stages 1 through $b-1$ are unchanged. Hence, as the input rate λ increases, while the flow through all these stations increases proportinally, resulting in higher production cost for the first part of the line (up to stage b).

On the other hand, if $l_\lambda < b$, let $U'_R(\lambda)$ be the set of stages $i < b$ with a bottleneck repair station for input rate λ. Then by Proposition A.0.1 and Assumption 4.4.2, we have that $U_R(\lambda_1) \subseteq U_R(\lambda)$, that the fraction of units having the defects in $D_{i'}$ for $i' \in U'_R(\lambda)$ is nondecreasing in λ_2 for stages $i > i'$, and the fraction of units having defects in $D \setminus \cup_{i \in U'_R(\lambda)} D_i$ are not affected. Thus, the change in defect fractions for $j \in D_{i'}$, where $i' \in U'_R(\lambda)$ is propagated until the end of the line, and not affected by any other stage. However, under Assumption 4.4.2, the routing probabilities out of all stations in stage $i' \notin U'_R(\lambda)$ are unaffected by the change in the fraction of units having defects in $D_{i'}$ for $i' \in U'_R(\lambda)$. This implies that increasing the flow through the stations $i < b$ results in higher production cost for the first part of the line (up to stage b) under Assumption 4.4.3.

Summarizing the effects on the first part of the line, we observe that the flow through all the stations up to and including the operation station O_b is nondecreasing in λ, along with the fraction of units having defects in D_i for $i \in U'_R(\lambda)$. The next step is to study the effects of increased input flow on the second part of the line, after the bth stage. Note that even though the input flow is increased, the outflow from operation station O_b is constant at μ_b^O, as well as the fraction of units having defects in D_i for $i \geq b$ (under Assumption 4.4.2). This means that for the second part of the line, there are no changes that would affect the flow of parts or costs (by Assumption 4.4.3), except for the fact that the fraction of units having defects in D_i for $i \in U'_R(\lambda)$ is nondecreasing through the end of the production line.

Hence by stabilizing operation station O_b, we can reduce the total cost incurred up to stage b, while improving product quality at stage b. For the second part of

the line, after the bottleneck station b, the production, inspection, and repair costs are not affected. By Assumption 4.4.3, reduced final product quality together with increased production cost for the first part of the line and unaffected revenue implies decreased profitability of the system. The same argument applies if the first unstable operation or inspection station is an inspection station. Repeating this process for all unstable operation and inspection stations shows that stabilizing all such stations improves the total profit rate for the system. □

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER 5

Proof of Theorem 5.2.1. Note that the objective function $R^*(\boldsymbol{n}, \boldsymbol{\varepsilon})$ in (141) is concave since $\boldsymbol{H}$ is positive definite. Moreover, the constraints $(142) - (146)$ are linear. Hence the first order KKT conditions are necessary and sufficient for optimality. Let λ_1, λ_2, and λ_3 be the Lagrangian multipliers associated with the constraints $(142) - (144)$, respectively. Note that we can ignore the nonnegativity constraints $(145) - (146)$, because the optimum quantities and prices will always be nonnegative. Expanding the objective function in (141) yields

$$-R(\boldsymbol{Q}) = \frac{1}{d}\bigg(- (\alpha_2\varepsilon_1 + \beta\varepsilon_2)Q_1 + \alpha_2 Q_1^2 + 2\beta Q_1 Q_2 - (\alpha_1\varepsilon_2 + \beta\varepsilon_1)Q_2 + \alpha_1 Q_2^2\bigg),$$

and the KKT conditions become

$$\begin{aligned}
-\mu_1(\frac{\alpha_2}{d}\varepsilon_1 + \frac{\beta}{d}\varepsilon_2) + 2\frac{\alpha_2}{d}\mu_1 Q_1^* + 2\frac{\beta}{d}\mu_1 Q_2^* + \lambda_1 + \lambda_3 &= 0; && (163)\\
-\mu_2(\frac{\alpha_1}{d}\varepsilon_2 + \frac{\beta}{d}\varepsilon_1) + 2\frac{\alpha_1}{d}\mu_2 Q_2^* + 2\frac{\beta}{d}\mu_2 Q_1^* + \lambda_2 + \lambda_3 &= 0; && (164)\\
\lambda_1(\frac{Q_1^*}{\mu_1} - n_{1e}) &= 0; && (165)\\
\lambda_2(\frac{Q_2^*}{\mu_2} - n_{2e}) &= 0; && (166)\\
\lambda_3(\frac{Q_1^*}{\mu_1} + \frac{Q_2^*}{\mu_2} - n_e) &= 0; && (167)\\
\lambda_1, \lambda_2, \lambda_3 &\geq 0. && (168)
\end{aligned}$$

The first two conditions $(163) - (164)$ are obtained by taking the derivatives of $(141) - (144)$ with respect to Q_1 and Q_2, respectively, modified by the appropriate Lagrangian multipliers. The conditions $(165) - (167)$ are the complementary slackness conditions, and (168) provides the nonnegativity conditions for the Lagrangian multipliers.

Then we can solve equations $(163) - (164)$ for the optimum quantities in terms of λ_1, λ_2, and λ_3 as follows

$$\begin{aligned}
Q_1^* &= \frac{\varepsilon_1}{2} - \frac{\alpha_1}{2\mu_1}\lambda_1 - \frac{\gamma_2}{2\mu_1\mu_2}\lambda_3 + \frac{\beta}{2\mu_2}\lambda_2, && (169)\\
Q_2^* &= \frac{\varepsilon_2}{2} - \frac{\alpha_2}{2\mu_2}\lambda_2 - \frac{\gamma_1}{2\mu_1\mu_2}\lambda_3 + \frac{\beta}{2\mu_1}\lambda_1. && (170)
\end{aligned}$$

Equation (134) now yields

$$\begin{aligned} P_1^* &= \frac{\mu_1(\frac{\alpha_2}{d}\varepsilon_1 + \frac{\beta}{d}\varepsilon_2) + \lambda_1 + \lambda_3}{2\mu_1}, \\ P_2^* &= \frac{\mu_2(\frac{\alpha_1}{d}\varepsilon_2 + \frac{\beta}{d}\varepsilon_1) + \lambda_2 + \lambda_3}{2\mu_2}. \end{aligned}$$

Note that $P_i^* \geq 0$ for $i = 1, 2$ since $d > 0$. To see that Q_i^*, for $i = 1, 2$, can not be negative, assume that $Q_i^* < 0$. Then $Q_i^* < \mu_i n_{ie}$ and $Q_{3-i}^* \leq \mu_{3-i} n_{(3-i)e}$, implying that $Q_i^*/\mu_i + Q_{3-i}^*/\mu_{3-i} < n_{(3-i)e} \leq n_e$, and hence $\lambda_i = \lambda_3 = 0$. It now follows from equations (169) and (170) that $Q_i^* \geq 0$, a contradiction. In the optimum solution, the constraints (142) − (144) can be binding or non-binding with the corresponding multipliers nonnegative and zero, respectively. We analyze each case to construct different optimality scenarios.

First assume that all constraints are non-binding at the optimal solution, so that $\lambda_i = 0$ for all i. Then solving (169) − (170) for Q_1^* and Q_2^*, we obtain $Q_i^* = \varepsilon_i/2$. The production quantities Q_1 and Q_2 also need to satisfy the primary constraints (142) − (144) strictly, so that

$$\boldsymbol{\varepsilon} \geq 0; \varepsilon_1 < 2\mu_1 n_{1e}; \varepsilon_2 < 2\mu_2 n_{2e}; \frac{\varepsilon_1}{2\mu_1} + \frac{\varepsilon_2}{2\mu_2} < n_e,$$

corresponding to the solution in $\Omega_1(\boldsymbol{n})$.

Secondly, consider the case where (142) and (144) are non-binding and (143) is binding, so that $\lambda_1 = \lambda_3 = 0$ and $\lambda_2 \geq 0$. Then solving (169)−(170) for Q_1^* and λ_2 with $Q_2^* = \mu_2 n_{2e}$, we obtain $Q_1^* = \varepsilon_1/2 + \beta(\varepsilon_2/2 - \mu_2 n_{2e})/\alpha_2$ and $\lambda_2 = \mu_2(\varepsilon_2 - 2\mu_2 n_{2e}))/\alpha_2$. We also need to satisfy the primary constraints (142) and (144) strictly, so that $Q_1^* < \mu_1 n_1$, and ensure the nonnegativity of $\lambda_2 \geq 0$. These two conditions translate into $\Omega_2(\boldsymbol{n})$.

Thirdly, consider the case where (143) and (144) are non-binding and (142) is binding, so that $\lambda_2 = \lambda_3 = 0$ and $\lambda_1 \geq 0$. Then solving (169)−(170) for Q_2^* and λ_1 with $Q_1^* = \mu_1 n_{1e}$, we obtain $Q_2^* = \varepsilon_2/2 + \beta(\varepsilon_1/2 - \mu_1 n_{1e})/\alpha_1$ and $\lambda_1 = \mu_1(\varepsilon_1 - 2\mu_1 n_{1e})/\alpha_1$. We also need to satisfy the primary constraints (143) and (144) strictly, so that $Q_2^* < \mu_2 n_2$, and ensure the nonnegativity of $\lambda_1 \geq 0$. These two conditions translate into $\Omega_3(\boldsymbol{n})$.

Now consider the case where $n_f = 0$. Then (144) is binding if and only if (142) and (143) are binding. Thus (144) is not needed, and we can assume $\lambda_3 = 0$. It remains to consider the case where (142) and (143) are binding. Equations (163) and (164) with $Q_1^* = \mu_1 n_1$, $Q_2^* = \mu_2 n_2$, $\lambda_1 \geq 0$, and $\lambda_2 \geq 0$ yield that this solution is

optimal on

$$\Omega_{45}(\boldsymbol{n}) = \Big\{\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} \geq 0; \varepsilon_2 + \frac{\beta}{\alpha_1}\varepsilon_1 \geq 2\mu_2 n_2 + \frac{\beta}{\alpha_1}2\mu_1 n_1; \varepsilon_1 + \frac{\beta}{\alpha_2}\varepsilon_2 \geq 2\mu_1 n_1 + \frac{\beta}{\alpha_2}2\mu_2 n_2\Big\}.$$

Note that $\Omega_6(\boldsymbol{n}) = \emptyset$ when $n_f = 0$ and that $\Omega_{45}(\boldsymbol{n})$ is the intersection of two half-planes defined by lines with negative slopes $-\beta/\alpha_1$ and $-\alpha_2/\beta$, respectively, where

$$-\frac{\beta}{\alpha_1} - \Big(-\frac{\alpha_2}{\beta}\Big) \;=\; \frac{d}{\alpha_1\beta} > 0.$$

It remains to show that $\Omega_{45}(\boldsymbol{n}) = \Omega_4(\boldsymbol{n}) \cup \Omega_5(\boldsymbol{n})$. Let $\Omega_0(\boldsymbol{n}) = \{\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} \geq 0; \varepsilon_1\gamma_1 - \varepsilon_2\gamma_2 \geq 2\mu_1 n_1\gamma_1 - 2\mu_2 n_2\gamma_2\}$ and $\Omega_0'(\boldsymbol{n}) = \{\boldsymbol{\varepsilon} : \boldsymbol{\varepsilon} \geq 0; \varepsilon_1\gamma_1 - \varepsilon_2\gamma_2 \leq 2\mu_1 n_1\gamma_1 - 2\mu_2 n_2\gamma_2\}$ be half-planes defined by a line with slope γ_1/γ_2, where

$$\frac{\gamma_1}{\gamma_2} - \Big(-\frac{\beta}{\alpha_1}\Big) \;=\; \frac{\mu_1 d}{\alpha_1\gamma_2}, \tag{171}$$

$$\frac{\gamma_1}{\gamma_2} - \Big(-\frac{\alpha_2}{\beta}\Big) \;=\; \frac{\mu_2 d}{\beta\gamma_2}. \tag{172}$$

The lines in the definitions of $\Omega_{45}(\boldsymbol{n})$, $\Omega_0(\boldsymbol{n})$, and $\Omega_0'(\boldsymbol{n})$ intersect at the point $(2\mu_1 n_1, 2\mu_2 n_2)$. It follows that $\Omega_4(\boldsymbol{n}) = \Omega_{45}(\boldsymbol{n}) \cap \Omega_0(\boldsymbol{n})$ and $\Omega_5(\boldsymbol{n}) = \Omega_{45}(\boldsymbol{n}) \cap \Omega_0'(\boldsymbol{n})$, both when $\beta \geq 0$ and $\gamma_2 > 0$ and when $\beta > 0$ and $\gamma_2 \leq 0$ (note that the case $\beta = 0$ and $\gamma_2 \leq 0$ is not possible). This proves the optimality of the specified solution on $\Omega_4(\boldsymbol{n})$ and $\Omega_5(\boldsymbol{n})$.

For the remainder of the proof, we assume that $n_f > 0$. Then (142) and (143) cannot be simultaneously binding. Next, consider the case where (142) and (144) are binding and (143) is non-binding, so that $\lambda_1, \lambda_3 \geq 0$ and $\lambda_2 = 0$, and hence $Q_1^* = \mu_1 n_{1e}$ and $Q_2^* = \mu_2 n_2$, meaning that all flexible capacity is assigned to product 1 and all of the dedicated capacity for product 2 is also used. Equations $(169) - (170)$ with $Q_1^* = \mu_1 n_{1e}$, $Q_2^* = \mu_2 n_2$, and $\lambda_2 = 0$ now yield

$$\begin{aligned} -2\mu_1 n_{1e}\gamma_1 &= -\gamma_1\varepsilon_1 + \frac{\gamma_1\alpha_1}{\mu_1}\lambda_1 + \frac{\gamma_1\gamma_2}{\mu_1\mu_2}\lambda_3, \\ 2\mu_2 n_2\gamma_2 &= \gamma_2\varepsilon_2 - \frac{\gamma_1\gamma_2}{\mu_1\mu_2}\lambda_3 + \frac{\beta\gamma_2}{\mu_1}\lambda_1, \end{aligned}$$

so that

$$\lambda_1 = \frac{\mu_1(2\mu_2 n_2\gamma_2 - 2\mu_1 n_{1e}\gamma_1 - \gamma_2\varepsilon_2 + \gamma_1\varepsilon_1)}{\beta\gamma_2 + \alpha_1\gamma_1}.$$

Also, it follows from (164) that

$$\lambda_3 = \frac{\mu_2}{d}(\alpha_1\varepsilon_2 + \beta\varepsilon_1 - 2\alpha_1\mu_2 n_2 - 2\beta\mu_1 n_{1e}).$$

Note that $\beta\gamma_2 + \alpha_1\gamma_1 = \mu_1(\alpha_1\alpha_2 - \beta^2) = \mu_1 d > 0$. The nonnegativity of λ_1 and λ_3 conditions for optimality translate into $\Omega_4(\boldsymbol{n})$.

Next, consider the case where (143) and (144) are binding and (142) is non-binding, so that $\lambda_2, \lambda_3 \geq 0$ and $\lambda_1 = 0$, and hence $Q_1^* = \mu_1 n_1$ and $Q_2^* = \mu_2 n_{2e}$, meaning that all flexible capacity is assigned to product 2 and all of the dedicated capacity for product 1 is used. Then solving (169) − (170) for λ_2 with $Q_1^* = \mu_1 n_1$, $Q_2^* = \mu_2 n_{2e}$, and $\lambda_1 = 0$, we obtain

$$\lambda_2 = \frac{\mu_2(2\mu_1 n_1\gamma_1 - 2\mu_2 n_{2e}\gamma_2 - \gamma_1\varepsilon_1 + \gamma_2\varepsilon_2)}{\beta\gamma_1 + \alpha_2\gamma_2}.$$

It follows from (163) that

$$\lambda_3 = \frac{\mu_1}{d}(\alpha_2\varepsilon_1 + \beta\varepsilon_2 - 2\alpha_2\mu_1 n_1 - 2\beta\mu_2 n_{2e}).$$

Note that $\beta\gamma_1 + \alpha_2\gamma_2 = \mu_2 d > 0$. The nonnegativity conditions for λ_2 and λ_3 translate into $\Omega_5(\boldsymbol{n})$.

Next, consider the case where (142) and (143) are non-binding and (144) is binding, so that $\lambda_1 = \lambda_2 = 0$ and $\lambda_3 \geq 0$, meaning that all the flexible capacity is shared between the two product groups. Let $a > 0$ be the amount of flexible capacity used for product 1, so that the remaining flexible capacity $fn_f - a > 0$ is reserved for product 2. Equations (169) − (170) with $Q_1^* = \mu_1(n_1 + a)$, $Q_2^* = \mu_2(n_{2e} - a)$, and $\lambda_1 = \lambda_2 = 0$ yield

$$\begin{aligned} 2\gamma_1\mu_1(n_1 + a) &= \gamma_1\varepsilon_1 - \frac{\gamma_1\gamma_2}{\mu_1\mu_2}\lambda_3, \\ -2\gamma_2\mu_2(n_{2e} - a) &= -\gamma_2\varepsilon_2 + \frac{\gamma_1\gamma_2}{\mu_1\mu_2}\lambda_3, \end{aligned} \tag{173}$$

so that

$$a = \frac{\varepsilon_1\gamma_1 - \varepsilon_2\gamma_2 + 2\mu_2 n_{2e}\gamma_2 - 2\mu_1 n_1\gamma_1}{2(\mu_1\gamma_1 + \mu_2\gamma_2)}.$$

This yields

$$\begin{aligned} Q_1^* &= \mu_1(n_1 + a) = \frac{\mu_1(2n_e\mu_2\gamma_2 - \varepsilon_2\gamma_2 + \varepsilon_1\gamma_1)}{2(\mu_1\gamma_1 + \mu_2\gamma_2)}, \\ Q_2^* &= \mu_2(n_{2e} - a) = \frac{\mu_2(2n_e\mu_1\gamma_1 + \varepsilon_2\gamma_2 - \varepsilon_1\gamma_1)}{2(\mu_1\gamma_1 + \mu_2\gamma_2)}. \end{aligned}$$

Moreover, it follows from (173) that

$$\lambda_3 = \frac{\mu_1\mu_2(\varepsilon_1\mu_2 + \varepsilon_2\mu_1 - 2\mu_1\mu_2 n_e)}{\mu_1\gamma_1 + \mu_2\gamma_2}.$$

To see that the denominator for a and λ_3 above can not be negative, note that the function $\mu_1\gamma_1 + \mu_2\gamma_2 = \mu_1(\alpha_2\mu_1 - \beta\mu_2) + \mu_2(\alpha_1\mu_2 - \beta\mu_1)$ is minimized with respect to μ_2 when $\mu_2 = \beta\mu_1/\alpha_1$. Hence, $\mu_1\gamma_1 + \mu_2\gamma_2 \geq \mu_1^2 d/\alpha_1 > 0$. The optimality conditions $a > 0$, $a < fn_f$, and $\lambda_3 \geq 0$ define the region corresponding to $\Omega_6(\boldsymbol{n})$. □

Proof of Proposition 5.2.1. The result follows from the definition of the regions $\Omega_i(\boldsymbol{n})$ for $i = 1, \ldots, 6$. It is easy to see that regions 1, 2, and 3 are not empty. To see that region 6 is also not empty, first note that the planar strip

$$-2\mu_2 n_{2e}\gamma_2 + 2\mu_1 n_1\gamma_1 < \varepsilon_1\gamma_1 - \varepsilon_2\gamma_2 < 2\mu_1 n_{1e}\gamma_1 - 2\mu_2 n_2\gamma_2$$

is always nonempty because

$$2\mu_1 n_{1e}\gamma_1 - 2\mu_2 n_2\gamma_2 - (-2\mu_2 n_{2e}\gamma_2 + 2\mu_1 n_1\gamma_1) = 2fn_f(\mu_1\gamma_1 + \mu_2\gamma_2) > 0,$$

see the last paragraph of the proof of Theorem 5.2.1. Furthermore, the line $\varepsilon_1/2\mu_1 + \varepsilon_2/2\mu_2 = n_e$ intersects the boundaries of the above nonempty strip at the points $(2n_1\mu_1, 2n_{2e}\mu_2)$ and $(2n_{1e}\mu_1, 2n_2\mu_2)$ in the positive quadrant, respectively, implying that $\Omega_6(\boldsymbol{n})$ is not empty.

Finally, note that the two lines defining each of regions $\Omega_4(\boldsymbol{n})$ and $\Omega_5(\boldsymbol{n})$ intersect in the positive quadrant at the points $(2n_{1e}\mu_1, 2n_2\mu_2)$ and $(2n_1\mu_1, 2n_{2e}\mu_2)$, respectively. The lines defining $\Omega_4(\boldsymbol{n})$ have slopes $-\beta/\alpha_1$ and γ_1/γ_2, and the lines defining $\Omega_5(\boldsymbol{n})$ have slopes $-\alpha_2/\beta$ and γ_1/γ_2. It now follows from (171) and (172) that $\Omega_4(\boldsymbol{n})$ and $\Omega_5(\boldsymbol{n})$ are non-empty in all possible cases (i.e., when $\beta \geq 0$ and $\gamma_2 > 0$ and when $\beta > 0$ and $\gamma_2 \leq 0$). □

Proof of Theorem 5.3.1. Similar to the proof of Theorem 5.2.1, this result follows by analyzing the first order KKT conditions for the nonlinear programming problem (149) − (155). More specifically, let λ_1^s, λ_2^s, and λ_3^s, $s = 1, \ldots, S$, be the Lagrangian multipliers associated with the constraints (150)−(152), respectively. Similarly, let u_1, u_2, and u_3 be the Lagrangian multipliers associated with the nonnegativity constraints in (153) for n_1, n_2, and n_f, respectively. We ignore the nonnegativity constraints (154) − (155), because the optimum quantities and prices will always be nonnegative (see below). As in the proof of Theorem 5.2.1, since the objective function $V^*(\boldsymbol{n}, \boldsymbol{\varepsilon})$ in (149) is concave and the set of constraints (150) − (155) are linear, the first order KKT conditions are necessary and sufficient for optimality, implying that any optimal solution should satisfy the primary conditions (150) − (155) as well as the associated

KKT conditions. Expanding the objective function in (149) yields

$$\begin{aligned} -V(\boldsymbol{Q},\boldsymbol{n}) &= \sum_{s\in S}\frac{r_s}{d}\Big(-(\alpha_2\varepsilon_1^s+\beta\varepsilon_2^s)Q_1^s+\alpha_2(Q_1^s)^2+2\beta Q_1^sQ_2^s-(\alpha_1\varepsilon_2^s+\beta\varepsilon_1^s)Q_2^s+\alpha_1(Q_2^s)^2\Big) \\ &\quad +c_1n_1+c_2n_2+c_fn_f, \end{aligned}$$

and the KKT conditions become

$$-r_s\mu_1(\frac{\alpha_2}{d}\varepsilon_1^s+\frac{\beta}{d}\varepsilon_2^s)+2r_s\frac{\alpha_2}{d}\mu_1Q_1^s+2r_s\frac{\beta}{d}\mu_1Q_2^s+\lambda_1^s+\lambda_3^s = 0,\forall s; \quad (174)$$

$$-r_s\mu_2(\frac{\alpha_1}{d}\varepsilon_2^s+\frac{\beta}{d}\varepsilon_1^s)+2r_s\frac{\alpha_1}{d}\mu_2Q_2^s+2r_s\frac{\beta}{d}\mu_2Q_1^s+\lambda_2^s+\lambda_3^s = 0,\forall s; \quad (175)$$

$$u_1+\sum_{s=1}^{S}(\lambda_1^s+\lambda_3^s) = c_1; \quad (176)$$

$$u_2+\sum_{s=1}^{S}(\lambda_2^s+\lambda_3^s) = c_2; \quad (177)$$

$$u_3+f\sum_{s=1}^{S}(\lambda_1^s+\lambda_2^s+\lambda_3^s) = c_f; \quad (178)$$

$$\lambda_1^s(\frac{Q_1^s}{\mu_1}-n_{1e}) = 0,\forall s; \quad (179)$$

$$\lambda_2^s(\frac{Q_2^s}{\mu_2}-n_{2e}) = 0,\forall s; \quad (180)$$

$$\lambda_3^s(\frac{Q_1^s}{\mu_1}+\frac{Q_2^s}{\mu_2}-n_e) = 0,\forall s; \quad (181)$$

$$u_1n_1=0,u_2n_2=0,u_3n_f = 0; \quad (182)$$

$$u_1,u_2,u_3,\lambda_1^s,\lambda_2^s,\lambda_3^s \geq 0,\forall s. \quad (183)$$

Then the optimum quantities Q_1^s, Q_2^s and prices P_1^s, P_2^s for each scenario s can be found as in the proof of Theorem 5.2.1, and are given by

$$\begin{aligned} Q_1^s &= \frac{\varepsilon_1^s}{2}-\frac{\alpha_1}{2r_s\mu_1}\lambda_1^s-\frac{\gamma_2}{2r_s\mu_1\mu_2}\lambda_3^s+\frac{\beta}{2r_s\mu_2}\lambda_2^s \\ &= \frac{\varepsilon_1^s}{2}-\frac{\alpha_1}{2r_s\mu_1}(\lambda_1^s+\lambda_3^s)+\frac{\beta}{2r_s\mu_2}(\lambda_2^s+\lambda_3^s), \quad (184) \\ Q_2^s &= \frac{\varepsilon_2^s}{2}-\frac{\alpha_2}{2r_s\mu_2}\lambda_2^s-\frac{\gamma_1}{2r_s\mu_1\mu_2}\lambda_3^s+\frac{\beta}{2r_s\mu_1}\lambda_1^s \\ &= \frac{\varepsilon_2^s}{2}-\frac{\alpha_2}{2r_s\mu_2}(\lambda_2^s+\lambda_3^s)+\frac{\beta}{2r_s\mu_1}(\lambda_1^s+\lambda_3^s), \quad (185) \\ P_1^s &= \frac{r_s\mu_1(\frac{\alpha_2}{d}\varepsilon_1^s+\frac{\beta}{d}\varepsilon_2^s)+\lambda_1^s+\lambda_3^s}{2r_s\mu_1}, \quad (186) \\ P_2^s &= \frac{r_s\mu_2(\frac{\alpha_1}{d}\varepsilon_2^s+\frac{\beta}{d}\varepsilon_1^s)+\lambda_2^s+\lambda_3^s}{2r_s\mu_2}. \quad (187) \end{aligned}$$

Note that the prices P_1^s, P_2^s in each scenario s are always positive. By a similar argument as in the proof of Theorem 5.2.1, the quantities Q_1^s and Q_2^s for each scenario s are also positive. Then, using equations (176)−(177) and (184)−(187), the expected optimal quantities and prices can be calculated as

$$\begin{aligned} E[Q_i^*] &= \sum_{s=1}^{S} Q_i^s r_s = \frac{1}{2}\sum_{s=1}^{S} r_s \varepsilon_i^s - \frac{\alpha_i}{2\mu_i}\sum_{s=1}^{S}(\lambda_i^s + \lambda_3^s) + \frac{\beta}{2\mu_j}\sum_{s=1}^{S}(\lambda_j^s + \lambda_3^s) \\ &= \frac{E[\xi_i]}{2} - \frac{\alpha_i(c_i - u_i)}{2\mu_i} + \frac{\beta(c_j - u_j)}{2\mu_j} \text{ for } i = 1, 2, j = 3 - i, \qquad (188) \\ E[P_i^*] &= \sum_{s=1}^{S} P_i^s r_s = \frac{1}{2d}\sum_{s=1}^{S} r_s(\alpha_j \varepsilon_i^s + \beta \varepsilon_j^s) + \frac{1}{2\mu_i}\sum_{s=1}^{S}(\lambda_i^s + \lambda_3^s) \\ &= \frac{E[\alpha_j \xi_i + \beta \xi_j]}{2d} + \frac{c_i - u_i}{2\mu_i} \text{ for } i = 1, 2, j = 3 - i. \qquad (189) \end{aligned}$$

(a) If the firm chooses to invest only in product i, then $u_i = 0$ and it is obvious that $Q_j^s = 0$ for $j \neq i$ and all s. Hence $E[Q_j^*] = 0$, implying that $(c_j - u_j)/\mu_j = (E[\xi_j]\mu_i + \beta c_i)/(\mu_i \alpha_j)$. Then equations (156) − (157) follow from equations (188) − (189) by letting $u_i = 0$.

(b) For contradiction, assume that the firm will introduce both products without investing in both of the dedicated capacities. First, consider the case where $n_1^*, n_f^* > 0$ and $n_2^* = 0$. By (182), this implies that $u_1 = u_3 = 0$, and the set of constraints (150) is redundant, implying that $\lambda_1^s = 0$ for all s. Then from equations (177) − (178), we have $c_2 = c_f/f + u_2$. However, it is impossible to satisfy the last equality with $u_2 \geq 0$, since by assumption $c_f/f > c_2$. By similar reasoning, the case where $n_2^*, n_f^* > 0$ and $n_1^* = 0$ is also impossible, because it implies $c_1 = c_f/f + u_1$. Finally, consider the case where $n_1^* = n_2^* = 0$ and $n_f^* > 0$, so that $u_3 = 0$, the constraints (150) − (151) are redundant, and $\lambda_1^s = \lambda_2^s = 0$. This implies that $c_1 = c_f/f + u_1$ and $c_2 = c_f/f + u_2$, a contradiction. Therefore, $n_1^*, n_2^* > 0$ if the firm produces both products, and hence $u_1 = u_2 = 0$. Then equations (158) − (159) follow from equations (188) − (189) by letting $u_1 = u_2 = 0$.

(c) The expected optimal production quantities clearly satisfy $E[Q_i^*] \geq 0$ for $i = 1, 2$. First consider the case where it is optimal to invest in both products. Then by (158), we have

$$\begin{aligned} E[\xi_1]\mu_1\mu_2 + \beta c_2 \mu_1 - \alpha_1 c_1 \mu_2 \geq 0, \\ E[\xi_2]\mu_1\mu_2 + \beta c_1 \mu_2 - \alpha_2 c_2 \mu_1 \geq 0. \end{aligned}$$

Multiplying the first inequality by α_2, the second one by β, and summing, we obtain $c_1 \leq \mu_1 E[\alpha_2 \xi_1 + \beta \xi_2]/d$. Similarly, multiplying the first inequality by β, the second one by α_1, and summing, we obtain $c_2 \leq \mu_2 E[\alpha_1 \xi_2 + \beta \xi_1]/d$. This shows that $c_i > \mu_i E[\alpha_j \xi_i + \beta \xi_j]/d$ implies that either $n_i^* = 0$ or $n_j^* = 0$, where $j = 3 - i$. Suppose now that $n_i^* > 0$ and $n_j^* = 0$. Then (156) implies that $E[Q_i^*] < 0$, a contradiction. Hence $c_i > \mu_i E[\alpha_j \xi_i + \beta \xi_j]/d$ implies that $n_i^* = 0$.

□

REFERENCES

[1] Ahn, H.-S., Duenyas, I., and Lewis, M. E., "The optimal control of a two-stage tandem queueing system with flexible servers," *Probability in the Engineering and Informational Sciences*, vol. 16, pp. 453–469, 2002.

[2] Ahn, H.-S., Duenyas, I., and Zhang, R., "Optimal stochastic scheduling of a two-stage tandem queue with parallel servers," *Advances in Applied Probability*, vol. 16, pp. 453–469, 1999.

[3] Ahn, H.-S., Duenyas, I., and Zhang, R., "Optimal control of a flexible server," *Advances in Applied Probability*, vol. 36, pp. 139–170, 2004.

[4] Andradóttir, S. and Ayhan, H., "Throughput maximization for tandem lines with two stations and flexible servers," *Operations Research*, vol. 53, pp. 516–531, 2005.

[5] Andradóttir, S., Ayhan, H., and Down, D. G., "Server assignment policies for maximizing the steady-state throughput of finite queueing systems," *Management Science*, vol. 47, pp. 1421–1439, 2001.

[6] Andradóttir, S., Ayhan, H., and Down, D. G., "Dynamic server allocation for queueing network with flexible servers," *Operations Research*, vol. 51, pp. 952–968, 2003.

[7] Andradóttir, S., Ayhan, H., and Down, D. G., "Compensating for failures with flexible servers," *Operations Research*, vol. 55, pp. 753–768, 2007.

[8] Andradóttir, S., Ayhan, H., and Down, D. G., "Dynamic assignment of dedicated and flexible servers in tandem lines," *Probability in the Engineering and Informational Sciences*, vol. 21, pp. 497–538, 2007.

[9] Bai, D. S. and Yun, H. J., "Optimal allocation of inspection effort in a serial multi-stage production system," *Computers and Industrial Engineering*, vol. 30, no. 3, pp. 387–396, 1996.

[10] Bell, S. L. and Williams, R. J., "Dynamic scheduling of a system with two parallel servers in heavy traffic with complete resource pooling: Asymptotic optimality of a continuous review threshold policy," *Annals of Applied Probability*, vol. 11, pp. 608–649, 2001.

[11] Bell, S. L. and Williams, R. J., "Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy," *Electronic Journal of Probability*, vol. 10, pp. 1044–1115, 2005.

[12] Bertsimas, D. and Tsitsiklis, J. N., *Introduction to Linear Optimization*. Belmont, MA: Athena Scientific, second ed., 1997.

[13] Biller, S., Muriel, A., and Zhang, Y., "Impact of price postponement on capacity and flexibility investment decisions," *Production and Operations Management*, vol. 15, no. 2, pp. 198–214, 2006.

[14] Birge, J. R., Drogosz, K., and Duenyas, I., "Setting single period optimal capacity levels and prices for substitutable products," *International Journal of Flexible Manufacturing Systems*, vol. 10, pp. 407–430, 1998.

[15] Bish, E. K. and Hong, J. H., "Coordinating the resource investment decision for a two-market, price setting firm," *International Journal of Production Economics*, vol. 101, pp. 63–88, 2006.

[16] Bish, E. K., Liu, J., and Suwandechochai, R., "Optimal capacity, product substitution, linear demand models, and uncertainty," *The Engineering Economist*, vol. 54, pp. 109–151, 2009.

[17] Bish, E. K. and Suwandechochai, R., "Impact of demand substitution/complementarity and price/quantity postponement on the optimal capacity of the flexible resource," in *Proceedings of the INFORMS MSOM 2005 Conference*, 2005.

[18] Bish, E. K. and Wang, Q., "Optimal investment strategies for flexible resources considering pricing and correlated demands," *Operations Research*, vol. 52, pp. 954–964, 2004.

[19] Boudette, N. E., "Chrysler gains edge by giving new flexibility to its factories," *The Wall Street Journal*, vol. April 11, p. A1, 2006.

[20] Bramson, M. and Williams, R. J., "On dynamic scheduling of stochastic networks in heavy traffic and some new results for the workload process," in *Proceedings of the 39th IEEE Conference on Decision and Control*, pp. 516–521, 2000.

[21] Britney, R. R., "Optimal screening plans for nonserial production systems," *Management Science*, vol. 18, pp. 550–559, 1972.

[22] Chen, A., "An alternative dynamic programming approach to allocating inspection points in multistage production systems," *Quality Engineering*, vol. 11, no. 2, pp. 197–205, 1998.

[23] Chen, H. and Mandelbaum, A., "Discrete flow networks: Bottleneck analysis and fluid limit approximations," *Mathematics of Operations Research*, vol. 16, pp. 408–445, 1991.

[24] Chen, H. and Mandelbaum, A., "Stochastic discrete flow networks: Diffusion approximations and bottlenecks," *The Annals of Probability*, vol. 19, no. 4, pp. 1463–1519, 1991.

[25] Chen, H. and Yao, D., *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. New York, US: Springer, 2001.

[26] Chen, H. and Zhang, H., "Stability of multiclass queueing networks under priority service disciplines," *Operations Research*, vol. 48, no. 1, pp. 26–37, 2000.

[27] Chevalier, P. B. and Wein, L. M., "Inspection for circuit board assembly," *Management Science*, vol. 43, pp. 1198–1213, 1997.

[28] Chod, J. and Rudi, N., "Resource flexibility with responsive pricing," *Operations Research*, vol. 53, no. 3, pp. 532–548, 2005.

[29] Cochran, J. K. and Erol, R., "Performance modeling of serial production lines with inspection/repair stations," *International Journal of Production Research*, vol. 39, pp. 1707–1720, 2001.

[30] Dai, J. G., "On positive Harris recurrance of multiclass queueing networks: A unified approach via fluid limit models," *Annals of Applied Probability*, vol. 5, pp. 49–77, 1995.

[31] Dai, J. G., "A fluid limit model criterion for instability of multiclass queueing networks," *Annals of Applied Probability*, vol. 6, pp. 751–757, 1996.

[32] Dai, J. G., *Stability of Fluid and Stochastic Processing Networks*, vol. 9 of *MaPhySto Miscellanea Publications*. Ny Munkegade, Denmark: Centre for Mathematical Physics and Stochastics, 1999.

[33] Dai, J. G. and Dai, W., "A heavy traffic limit theorem for a class of open queueing networks with finite buffers," *Queueing Systems*, vol. 32, pp. 5–40, 1999.

[34] Davis, M. H. A., "Piecewise deterministic Markov processes: A general class of diffusion stochastic models," *Journal of Royal Statistics Society: Series B*, no. 46, pp. 353–388, 1984.

[35] Edmondson, G., "BMW keeps the home fires burning," *Business Week*, March 30, 2005.

[36] Egorova, R., Borst, S., and Zwart, B., "Bandwidth-sharing networks in overload," *Performance Evaluation*, vol. 64, no. 9-12, pp. 978–993, 2007.

[37] Farrar, T. M., "Optimal use of an extra server in a two station tandem queueing network," *IEEE Transactions on Automatic Control*, vol. 38, pp. 1296–1299, 1993.

[38] Fine, C. and Freund, R., "Optimal investment in product flexible manufacturing capacity," *Management Science*, vol. 36, pp. 449–466, 1990.

[39] Fleischmann, B. S. and Henrich, P., "Strategic planning of BMW's global production network," *Interfaces*, vol. 36, no. 3, pp. 194–208, 2006.

[40] Foster, J. W., Malave, C. O., and Villalobos, J. R., "Flexible inspection within an aggregated information environment," *Computers and Industrial Engineering*, vol. 19, pp. 224–228, 1990.

[41] Galante, G. and Passannanti, G., "Integrated approach to part scheduling and inspection policies for a job shop manufacturing system," *International Journal of Production Research*, vol. 45, no. 22, pp. 5177–5198, 2007.

[42] Garcia-Diaz, A., Foster, J. W., and Bonyuet, M., "Dynamic programming analysis of special multi-stage inspection systems," *IIE Transactions*, vol. 16, pp. 115–125, 1984.

[43] Goodman, J. B. and Massey, W. A., "The non-ergodic Jackson network," *Journal of Applied Probability*, vol. 21, pp. 860–869, 1984.

[44] Goyal, M. and Netessine, S., "Capacity investment and the interplay between volume flexibility and product flexibility," in *MSOM Meeting, Northwestern University*, 2005.

[45] Goyal, M. and Netessine, S., "Strategic technology choice and capacity investment under demand uncertainty," *Management Science*, vol. 53, no. 2, pp. 192–207, 2007.

[46] Goyal, M. and Netessine, S., "Volume flexibility, product flexibility or both: The role of demand correlation and product substitution," *Working Paper*, 2010.

[47] Goyal, M., Netessine, S., and Randall, T., "Deployment of manufacturing flexibility: An empirical analysis of the North American automotive industry," *Working Paper, http://www.netessine.com/*, 2010.

[48] Gupta, D., Gerchak, Y., and Buzacott, A., "The optimal mix of flexible and dedicated manufacturing capacities: Hedging against demand uncertainty," *International Journal of Production Economics*, vol. 28, no. 3, pp. 309–320, 1992.

[49] Gurnani, H., Drezner, Z., and Akella, R., "Capacity planning under different inspection strategies," *European Journal of Operations Research*, vol. 89, pp. 302–312, 1989.

[50] Hajek, B., "Optimal control of interacting service stations," *IEEE Transactions on Automatic Control*, vol. 29, pp. 491–499, 1984.

[51] Han, M. S., Lim, J. T., and Park, D. J., "Performance analysis of serial production lines with quality inspection machines," *International Journal of Systems Science*, vol. 29, no. 9, pp. 939–951, 1998.

[52] Harrison, J. M. and López, M. J., "Heavy traffic resource pooling in parallel-server systems," *Queueing Systems*, vol. 33, pp. 339–368, 1999.

[53] Holweg, M. and Pil, F. K., "The second century: Reconnecting customer and value chain through build to order; moving beyond mass and lean production in auto industry," *MIT Press*, 2004.

[54] Hsu, L. F. and Tapiero, C. S., "Quality control of the M/G/1 queue," *European Journal of Operations Research*, vol. 42, pp. 88–100, 1989.

[55] Hsu, L. F. and Tapiero, C. S., "Economic model for determining the optimal quality and process control policy in queue-like production system," *International Journal of Production Research*, vol. 28, pp. 1447–1457, 1990.

[56] Hurst, E. G., "Imperfect inspection in a multistage production process," *Management Science*, vol. 20, pp. 378–384, 1973.

[57] Joncкheere, M., van der Mei, R. D., and van der Weij, W., "Rate stability and output rates in queueing networks with shared resources," *Performance Evaluation*, vol. 67, no. 1, pp. 28–42, 2010.

[58] Jordan, W. and Graves, S. C., "Principles on the benefits of manufacturing process flexibility," *Management Science*, vol. 41, no. 4, pp. 577–594, 1995.

[59] Kakade, V., Jorge, F. V., and Smith, J. S., "An optimization model for selective inspection in serial manufacturing systems," *International Journal of Production Research*, vol. 42, no. 18, pp. 3891–3909, 2004.

[60] Kim, J. and Gershwin, S. B., "Integrated quality and quantity modeling of a production line," *OR Spectrum*, vol. 27, pp. 287–314, 2005.

[61] Kogan, K. and Raz, T., "Optimal allocation of inspection effort over a finite planing horizon," *IIE Transactions*, vol. 34, pp. 515–527, 2002.

[62] Kopzon, A., Nazarathy, Y., and Weiss, G., "A push—pull network with infinite supply of work," *Queueing Systems*, vol. 62, no. 1-2, pp. 75–111, 2009.

[63] Kouikoglou, V. S. and Phillis, Y. A., "Design of product specifications and control policies in a single-stage production system," *IIE Transactions*, vol. 34, pp. 591–600, 2002.

[64] Lawler, G. F., *Introduction to Stochastic Processes*. CRC Press, second ed., 2006.

[65] Lee, J. and Unnikrishnan, S., "Planing quality inspection operations in multistage manufacturing systems with inspection errors," *International Journal of Production Research*, vol. 36, pp. 141–155, 1998.

[66] Lindsay, G. F. and Bishop, A. B., "Allocation of screening inspection efforts-a dynamic programming approach," *Management Science*, vol. 10, pp. 342–352, 1964.

[67] Lus, B. and Muriel, A., "Measuring the impact of increased product substitution on pricing and capacity decisions under linear demand models," *Production and Operations Management*, vol. 18, pp. 95–113, 2009.

[68] Mackintosh, J., "Ford learns to bend with the wind," *Financial Times*, vol. February 14, 2003.

[69] Mandroli, S. S., Shrivastava, A., and Ding, Y., "A survey of inspection strategy and sensor distribution studies in discrete-part manufacturing processes," *IIE Transactions*, vol. 38, pp. 309–328, 2006.

[70] McMurray, S., "Ford's f-150: Have it your way," *Business 2.0*, vol. March, pp. 53–55, 2004.

[71] Muriel, A., Somasundaram, A., and Zhang, Y., "Impact of partial manufacturing flexibility on production variability," *Manufacturing and Service Operations Management*, vol. 8, no. 2, pp. 192–205, 2006.

[72] Murty, K. G., *Linear Complementarity, Linear and Nonlinear Programming*, vol. 3 of *Sigma Series in Applied Mathematics*. Berlin, Germany: Helderman Verlag, 1988.

[73] Narahari, Y. and Khan, L. M., "Modeling reentrant manufacturing systems with inspection stations," *Journal of Manufacturing Systems*, vol. 15, pp. 367–378, 1996.

[74] Nazarathy, Y. and Weiss, G., "Positive Harris recurrence and diffusion scale analysis of a push-pull queueing network," *Performance Evaluation*, vol. 67, no. 4, pp. 201 – 217, 2010.

[75] Netessine, S., Dobson, G., and Shumsky, R., "Flexible service capacity: Optimal investment and the impact of demand correlation," *Operations Research*, vol. 50, pp. 375–388, 2002.

[76] Pandelis, D. G. and Teneketzis, D., "Optimal multiserver stochastic scheduling of two interconnected priority queues," *Advances in Applied Probability*, vol. 26, pp. 258–279, 1994.

[77] Penn, M. and Raviv, T., "A polynomial time algorithm for solving a quality control station configuration problem," *Discrete Applied Mathematics*, vol. 156, pp. 412–419, 2008.

[78] Rau, H. and Chu, Y. H., "Inspection allocation planning with two types of workstations: WVD and WAD," *International Journal of Advanced Manufacturing Technologies*, vol. 25, pp. 947–953, 2005.

[79] Rau, H., Chu, Y. H., and Cho, K. H., "Layer modeling for the inspection allocation problem in re-entrant production systems," *International Journal of Production Research*, vol. 43, no. 17, pp. 3633–3655, 2005.

[80] Raz, T., "A survey of models for allocating inspection effort in multistage production systems," *Journal of Quality Technology*, vol. 18, pp. 239–247, 1986.

[81] Raz, T. and Kaspi, M., "Location and sequencing of imperfect inspection operations in serial multi-stage production systems," *International Journal of Production Research*, vol. 29, pp. 1645–1659, 1991.

[82] Rosberg, Z., Varaiya, P. P., and Walrand, J. C., "Optimal control of service in tandem queues," *IEEE Transactions on Automatic Control*, vol. 27-3, pp. 600–609, 1982.

[83] Shiau, Y. R., "Inspection resource assignment in a multistage manufacturing system with an inspection error model," *International Journal of Production Research*, vol. 40, pp. 1787–1806, 2002.

[84] Shiau, Y. R., "Inspection allocation planning for a multiple quality characteristic advanced manufacturing system," *International Journal of Advanced Manufacturing Technology*, vol. 22, pp. 633–640, 2003.

[85] Shin, W. S., Hart, S. M., and Lee, H. F., "Strategic allocation of inspection stations for flow assembly line: a hybrid procedure," *IIE Transactions*, vol. 27, pp. 707–715, 1995.

[86] Tapiero, C. S. and Hsu, L. F., "Quality control of the M/M/1 queue," *International Journal of Production Research*, vol. 25, pp. 447–453, 1987.

[87] Tassiulas, L. and Bhattacharya, L. L., "Allocation of interdependent resources for maximal throughput," *Stochastic Models*, vol. 16, pp. 27–48, 2000.

[88] Tassiulas, L. and Ephremides, A., "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Transactions on Automatic Control*, vol. 37, pp. 1936–1948, 1992.

[89] Tolinski, M., "Hands-off inspection," *Society of Manufacturing Engineers*, vol. 135, no. 3, 2005.

[90] Van Mieghem, J. A., "Investment strategies for flexible resources," *Management Science*, vol. 44, pp. 1071–1078, 1998.

[91] Van Mieghem, J. A., "Commonality strategies: Value drivers and equivalence with flexible capacity and inventory substitution," *Management Science*, vol. 50, pp. 419–424, 2004.

[92] Van Mieghem, J. A. and Dada, M., "Price versus production postponement: Capacity and competition," *Management Science*, vol. 45, pp. 1631–1649, 1999.

[93] Villalobos, J. R., Foster, J. W., and Disney, R. L., "Flexible inspection systems for serial multi-stage production systems," *IIE Transactions*, vol. 25, pp. 16–26, 1993.

[94] Volsem, S. V., Dullaert, W., and Landeghem, H. V., "An evolutionary algorithm and discrete event simulation for optimizing inspection strategies for multi-stage processes," *European Journal of Operations Research*, vol. 179, pp. 621–633, 2007.

[95] Weiss, G., "Jackson networks with unlimited supply of work," *Journal of Applied Probability*, vol. 42, no. 3, pp. 879–882, 2005.

[96] Wiel, S. A. V. and Vardeman, S. B., "A discussion of all-or-none inspection policies," *Technometrics*, vol. 36, pp. 102–109, 1994.

[97] Williams, R. J., "On dynamic scheduling of a parallel server system with complete resource pooling," *Analysis of Communication Networks: Call Centers, Traffic and Performance*, vol. 28, 2000.

[98] Wu, C.-H., Lewis, M. E., and Veatch, M., "Dynamic allocation of reconfigurable resources in a two-stage tandem queueing system with reliability considerations," *IEEE Transactions on Automatic Control*, vol. 51, pp. 309–314, 2006.

[99] Yum, B. J. and McDowell, E. D., "The optimal allocation of inspection effort in a class of nonserial production systems," *IIE Transactions*, vol. 13, pp. 285–293, 1981.